UNIVERSITY OF DELHI
SCIENTIA ET MORES

RATAN TATA
LIBRARY

DELHI SCHOOL OF ECONOMICS

THE RATAN TATA LIBRARY

Cl. No. B280x9J 2G7

Ac. No. 11910

Date of release for loan

This book should be returned on or before the date last stamped below. An overedue charge of one anna will be levied for each day the book is kept beyond that date.

~~RESERVED~~

9·XI·5

10 3 58

25/7/58

2/8/58

27 DEC 1958

1967

15 JUL 1976

~~10 JAN 1985~~

APPLICATION OF STATISTICAL METHODS TO AGRICULTURAL RESEARCH

BY

HARRY H. LOVE, PH.D.

FORMERLY AGRICULTURAL ADVISER TO THE MINISTRY OF INDUSTRIES OF THE NATIONAL GOVERNMENT OF THE REPUBLIC OF CHINA, AND DIRECTOR OF CROP IMPROVEMENT FOR THE NATIONAL AGRICULTURAL RESEARCH BUREAU; PROFESSOR OF PLANT BREEDING, CORNELL UNIVERSITY

PUBLISHED UNDER THE AUSPICES OF THE NATIONAL AGRICULTURAL RESEARCH BUREAU AND THE CHINA FOUNDATION FOR THE PROMOTION OF EDUCATION AND CULTURE

THE COMMERCIAL PRESS, LIMITED
SHANGHAI
1937

DEDICATION

THIS BOOK IS DEDICATED TO THOSE
STUDENTS IN CHINA WHO HAVE SHOWN SUCH
DEEP INTEREST IN THE SUBJECT AND WHO ARE TAKING
THE LEAD IN THE APPLICATION OF STATISTICAL
ANALYSIS IN THE INTERPRETATION
OF THEIR DATA

PREFACE

For some time the author has been engaged in teaching biometrical methods to students in America and China. During this time he has had opportunity also to examine and analyze data collected by investigators engaged in many different fields of research. These contacts have impressed him with the need of a text that would present in simple form the fundamentals of statistical analysis, including the application of the experimental error concept and the method of variance analysis, in such a way as to enable the beginner to follow through the various steps needed for the analysis of his data.

The author realizes that there are many excellent texts, both general and specific, available on the subject. However, there does not seem to be any simple text combining in one volume methods for the measurement of variation, correlation, curve fitting, the probable error concept, and the analysis of variance. It is for this purpose that the present volume is prepared, and emphasis is placed on the interpretation and application of formulas rather than on their derivation and development. While particular consideration is given to the application of the different methods to agricultural data and examples are drawn largely from this field, the methods can be applied to any data that can be treated statistically. On account of the increasing interest in the problems of plot technic on the part of many investigators, the final chapter is given to a brief discussion of some of the more important phases of this subject.

The preparation and publication of this text have been made possible by funds furnished coöperatively by the China Foundation for the Promotion of Education and Culture and the National Agricultural Research Bureau of the Ministry of Industries of the National Government of the Republic of China. The author expresses his deep appreciation to these two organizations for their generous aid and coöperation.

The text is being translated into Chinese by Li-ying Shen (Mrs. T. H. Shen) of the National Agricultural Research Bureau, and especial thanks are due her for her careful and painstaking work in this connection, as well as for her helpful suggestions.

Data included in the text have been obtained from various sources. Efforts have been made to give credit in all cases, and the author is grateful to those whose data have been used. He is particularly indebted to Professor George W. Snedecor for permission to include his tables, which appear as tables X and XI in the Appendix, and to Dr. J. R. Livermore for the use of table IX in the Appendix. Dr. J. Wishart read most of the manuscript, and thanks are due him for many valuable criticisms and suggestions.

Finally, the author is grateful to Miss Frances Feehan for the excellent assistance she has rendered in all phases of the work, and feels that without her help it would have been impossible to have prepared the text.

Efforts have been made to eliminate errors, but the author realizes how difficult this is and will appreciate having any corrections called to his attention.

H. H. LOVE

August, 1934

CONTENTS

PAGE
Preface v
CHAPTER
I. Introduction 1
II. Frequency Distributions 11
III. Graphic Illustration 27
IV. Constants of Position 39
V. Constants of Deviation or Dispersion 66
VI. Simple Correlation 90
VII. Simple Correlation—*Continued* 132
VIII. Multiple and Partial Correlation 173
IX. The Probable Error Concept 216
X. Curve Fitting 249
XI. Goodness of Fit 287
XII. Analysis of Small Samples and Application of Probability 297
XIII.✓ Analysis of Variance 340
XIV✓ Analysis of Variance—Complex Experiment . . 378
XV. Problems of Plot Technic 398

APPENDIX

TABLE PAGE
I. Sums of Powers of Natural Numbers 468
II. Tabled Values to Facilitate the Fitting of a Logarithmic Curve of the General Formula $y=a+bx+c \log x$ 469
III. Table Giving Corresponding Values for r Computed from r_r 472

TABLE PAGE

IV. VALUES FOR FACILITATING COMPUTATION OF THE PROBABLE ERROR OF A SINGLE OBSERVATION, AND OF THE MEAN, FROM BESSEL'S FORMULA . . . 473

V. VALUES FOR FACILITATING COMPUTATION OF THE PROBABLE ERROR OF A SINGLE OBSERVATION, AND OF THE MEAN, FROM PETERS' FORMULA 474

VI. TABLE FOR ESTIMATING PROBABILITY BASED ON THE NORMAL PROBABILITY INTEGRAL CORRESPONDING TO VALUES OF $\frac{x}{\sigma}$ 475

VII. TABLE GIVING ODDS FOR VARIOUS VALUES OF $\frac{D}{P.E.}$. 479

VIII. THE CALCULATED ODDS FOR THE Z VALUES OF 'STUDENT'S' TABLE FOR ESTIMATING THE PROBABILITY OF THE SIGNIFICANCE OF THE RESULT 480

IX. TABLE OF ODDS CALCULATED FROM 'STUDENT'S' t TABLE 484, 485

X. VALUES OF F AND t 486

XI. SIGNIFICANT VALUES OF r AND R 490

PUBLICATIONS REFERRED TO IN THE TEXT 493

INDEX 497

APPLICATION OF STATISTICAL METHODS TO AGRICULTURAL RESEARCH

CHAPTER I

INTRODUCTION

During recent years there has been a growing tendency in most fields of investigation to reduce the observations and data accumulated to an orderly arrangement that will make possible the evaluation of the results by means of some systematic method of analysis. That is, there has been a change from the earlier practices to more exact methods of expressing the results statistically. Formerly in a number of fields of investigation facts were observed and tabulated in such a way that it was rather difficult for anyone but the experimenter to gain a clear idea of the results. During these earlier years the field of statistics and statistical analysis had a few supporters, and the methods were gradually being developed and simplified so that some of the methods could be used by the non-mathematically trained investigator.

Formerly it was thought by many that statistical analysis belonged to a very specialized field, but more and more the methods of statistical analysis are finding their way into the workshops of scientists in all fields. This has been due largely to the fact that some of the enthusiastic supporters of the methods of statistics have worked faithfully to develop the methods so that they may be useful and usable for those persons who are not particularly trained in higher mathematics. Another reason was the real need felt on the part of students in biology, agriculture, sociology, education, psychology, economics, and the like, for some method that would enable them to systematize their records and reduce the facts which their results gave to a few simple statements. The result is that during recent years there has been a very rapid advancement in the field of statistical analysis and many useful methods are now available for the analysis of data arising from different sources. Many excellent textbooks, some of them general and some of them devoted to a special field of statistical analysis, have been written and are available to the public.

While there has been a very rapid growth in the application of statistics to the affairs of everyday life, in some cases mistakes have been made by those who are taking up some of the methods for the first time. It cannot be said that all mistakes are due to amateurs, but often a beginner is apt to feel that the application of some form of mathematics to the analysis of his data may make up or overcome all of the shortcomings or weaknesses that may be present in the data themselves. It must be understood at the outset that no conclusions, even if based on the most intricate mathematical analysis, can ever be more reliable than are the data which have been used in the analysis.

Collection of Data. This leads to the question of the kind and amount of data to collect, and we may say that the data should be based on a sufficiently large sample to be representative of the condition which it is intended to analyze. We should never form an opinion or come to any conclusion so that we may be biased in the collection of data, and we must be certain that the material studied is a fair sample. For example, if we were to study the health of the children in a certain community or city, we would not go to the best part of the city or to the best homes and take our data from the children found there. Neither would we collect our data from only the poorest part of the city, nor from the workshops. It would be necessary to study children from all environments of the city so that we could say the data were a fairly representative cross section of the children of that community.

Again, if we are to study the incomes of a group of merchants we would not go to a city and collect the data from only the merchants on the more prominent streets. It would be necessary also to have data from the small stores on the side streets if the data are to be fairly representative of the group as a whole. Again, if we are interested in studying the variation of a lot of wheat plants, it is not sufficient to go into a field and collect all of the plants in one corner of the field. The field will differ in its fertility and naturally the growth of the plants will be affected, and it is necessary to collect the plants so that they are truly representative of all of the plants that may be in the field. This leads to the term *random sample.*

Random Sample. What is meant by a *random sample?* It means that in collecting a sample, or a lot of individuals that will make up a sample, one must make the collection purely at random, that is without any bias whatsoever in the selection. It does not mean that one can discard individuals at will unless for some reason they do not represent normal reactions to a particular environment or condition from which they have been taken.

Referring again to the matter of collecting wheat plants for study, one can look at a field and see that a certain number of the plants are tall, another lot quite short, and the remainder of the plants about average, and then set out to collect plants on the basis of the estimates he has made. This will not lead to a random sample, since one cannot estimate the proper number of the different types to collect without exact counts. A random sample would be obtained if it were possible to go through a field of wheat, taking every tenth or twentieth plant in each row regardless of the conditions under which it has been produced, unless it is evident that it has been injured. Taking every tenth plant eliminates any bias or personal choice. One may have too large a sample for his study when he is finished, but it would again be possible to take another sample from this first collection by placing the plants together and again taking every tenth plant.

If one were interested in studying the height of the men in a certain city it would be necessary that his material be a random sample, and such a sample may be made up in any one of several ways. If it is a city having a directory, one could take every tenth male in the directory, or every fiftieth, depending on the size of the population and the number it is intended to study. Again, it would be possible to take the data on certain of the important corners where traffic is heavy, and study every tenth or twentieth individual, or whatever number may be decided on. If this were done at several important corners the result would be a random sample, providing the days on which the samples were taken were representative days. They should be business days and not holidays, since on holidays a certain class of people would not be appearing on the busy street corners, while those who do not ordinarily appear might be out in greater numbers. In order, then, that the data may be representative, it is necessary that it be collected at random, and based on sufficient observations to give a fair sample.

Measuring and Recording Data. After the data have been collected it is necessary to measure and tabulate the results. One should give careful consideration to the methods that are to be used and to the steps that are to be followed in connection with this work. In the first place, it is important that extreme accuracy be observed at all points. Even if one expects to apply the most involved methods of mathematical analysis to the data, it does not mean that such application will smooth out, eliminate, or reduce inaccuracies in measuring and recording the data.

One must make sure that the apparatus for measuring and weighing are correct so that they will give correct readings to the degree of exactness that it is intended to follow in making such measurements. If the rule or the scale is not reading properly, then each record will be influenced by an error. The errors may be proportional to the size of the item or they may not, depending on the actual discrepancies of the apparatus.

Since in most cases it is not possible to check all the measurements, readings, and observations made in collecting the data, it is of the greatest importance that this work be done with extreme accuracy. Such manipulations as divisions and multiplications made from the records may be checked and errors eliminated, but due to the fact that with most data one does not have an opportunity to recollect them, it is therefore impossible to check on the accuracy. The fact that the measurements cannot be checked should not be considered an excuse for careless work but an important argument for extreme accuracy.

In all cases wherever it is possible the measuring and recording should be checked. For example, if it is possible for two people to work together in weighing a lot of objects, one to do the weighing and one to do the recording, it is possible to check the actual taking of the weights and the recording. The person doing the weighing reads the weight to the person recording, and the record is made. Then the object is reweighed, letting the person doing the recording read back the weight to the person in charge of the weighing. This will slow up the work to some extent but will make for greater accuracy.

When the objects are being weighed on the type of balance that requires the changing or handling of several weights, it is better to devise some system that does not require the reading of the total amount of the weights on the scale pan. For example, with certain work the smaller types of laboratory balances are used and these may require the handling and rehandling of gram weights. If there is a great deal of weighing to be done it is more convenient to mark off a chart on a card that is placed directly in front of the balance, recording on this chart the various sized weights, as for example 500 grams, 200 grams, 100 grams, and so on. When these weights are not on the scale pan they should be placed on the chart over the proper number. Then when an object is weighed the weight may be quickly counted up by glancing at the chart and adding those numbers on the vacant places together with any weight that may be on the beam of the balance.

For checking the measuring and recording of data other devices may be developed, and it is worth while for each laboratory to consider ways and means of reducing errors or mistakes to a minimum. The recording of the figures must be carefully done, and the figures recorded so as to leave no doubt as to the numbers intended. For example, after recording numbers rapidly there may be doubt as to whether a 3 or a 5 is meant, and again as to whether a 7 or a 9 is meant.

The question of the fineness of the readings to be used when measuring or weighing objects needs to be considered, since it may be possible to save a considerable amount of time with certain kinds of data. For example, suppose that the weights of grain for several hundred wheat plants are to be obtained with the purpose of determining the average weight. If this is all the information desired from this weight value it is not necessary that time be spent reading the weighing to hundredths, or even tenths, of a gram. Sufficient information may usually be obtained by reading the weights to the whole unit, or gram.

With this kind of material it is often possible to save time by recording the weights directly into groups or classes. In Chapter II the arrangement of data in classes is fully discussed, and considerable time may be saved with such a problem as the weighing of grain by recording the individuals directly in the proper classes. In doing this it would not be necessary to know the exact weight of any of the individuals since the groups or classes would usually differ by at least a unit, as one gram, if not more, and therefore the group or class to which a particular individual belongs may be readily determined without noting its weight exactly. It would only be necessary to be more exact regarding those individuals that are near the division points between any two groups. This suggestion is made in some detail to call attention to the possibility of saving time in the weighing and recording of observations. There is no need to make the readings to a very fine point if later the objects are to be grouped in large classes. It may be, however, that there will be other facts wanted and then it will be desirable to make the readings to a finer degree, but this point should be determined if possible before much time is spent in detailed measurements.

Calculation of Results. In the calculation of results it is important that extreme care be observed. This means that for the more careful work systems for checking each operation should be devised. Where there is to be considerable statistical work done it is important that certain laboratory rules be established to be used in making and checking the calculations.

In the first place, for uniformity it is best to adopt certain standards as to the number of decimal places to be retained, and to adopt a rule for raising or lowering a number when reading to a certain number of decimal places. For example, if the number 4.235 has been obtained and the laboratory practice is to read to two decimals, a standard should be set as to what is to be done with the 5 in this case. Will the number in the second decimal place be raised and the reading made 4.24, or will the 5 be dropped? In some laboratories the rule is to always raise the number, while in others if the third decimal is exactly 5 it is dropped, and the number would be read 4.23. If it is 5+, the number is read 4.24. In certain operations it is possible to use the following practice. The first time a number comes out exactly 5 in the last decimal the number preceding may be raised, the second time the 5 may be dropped, and so on.

It is also important to adopt standards relative to the handling of squares and square roots. For example, if a certain division is to be made and the square root of the quotient extracted, and the answer is to be correct to two decimals, it is necessary that the quotient be carried to four decimal places. If three decimals are to be kept in the answer, then for accuracy the square or quotient should be carried to six decimal places. This is especially important when the square root is to be used in making multiplications and for other computations, and it is therefore very necessary for the greatest accuracy that a sufficient number of decimals be retained.

In connection with the matter of laboratory standards, it is important that certain operations always be handled in the same way and the same number of decimals retained. This is particularly true where many studies of a similar kind are being investigated and various comparisons are to be made between the different results.

Interpretation of Results. The most important part of any statistical analysis is the proper interpretation of results. It is a matter of common knowledge that the same statistics may sometimes be used to prove the very opposite result. This is often due to a selection of part of the statistics and to a neglect of some of the more important factors that may not have been considered. For example, Chaddock gives a good illustration of the danger of not considering all of the factors when drawing conclusions. His data are taken from a report of a college relative to the amount of smoking done by the students and the failures so far as their class work is concerned. The data as given by Chaddock are

	NON-SMOKERS	MODERATE SMOKERS	EXCESSIVE SMOKERS
NUMBER OF STUDENTS INVESTIGATED	111	35	18
AVERAGE WORK FOR YEAR	85.2%	73.3%	59.7%
PROPORTION OF FAILURES	3.2%	14.1%	24.1%

Taking the data as they are it is apparent that there seems to be some high relation between excessive smoking and failure in college work, but the question arises as to whether this is due to smoking or due to the fact that those who smoke a great deal may be those who are more interested in the social life of the college, or those who probably went to college not so much to obtain training as to pass a few years with pleasant associates. In other words, there are other factors that need to be considered, and one should not draw the conclusion that excessive smoking leads to failure in classes. On the other hand, it is true that the data do not warrant the conclusion that excessive smoking aids in college work.

One of the great dangers in statistical interpretation is due to the fact that perhaps inexperienced investigators, or those who may be using statistical methods for the first time, are apt to be imbued with the idea that the mathematical manipulations and the application of difficult formulas have a tendency to remove their data from the field of actuality, and therefore they fail to keep a proper balance between mathematical analysis and common sense. We should never lose sight of the common sense side of the problem, and should not be led to the drawing of conclusions that may not be justified by the affairs of everyday life even though mathematical manipulations seem to indicate otherwise. This does not mean that we should not announce a new or outstanding result which has been obtained from the application of statistical analysis, but we must study all of the factors involved and make certain that the result we are about to announce may be substantiated when the data are critically examined from all viewpoints.

There is one source of error in the handling and interpretation of statistics that is quite common, and one which affords numerous opportunities for incorrect conclusions or interpretations. This is the matter of the use of percentages. As an illustration we may consider the following data, which are the results obtained from two different methods of preparing soil for the growing of a crop of feterita.

A	B	GAIN OF B OVER A	GAIN IN PER CENT
4.7	9.8	5.1	108.51
7.1	12.8	5.7	80.28
25.0	28.4	3.4	13.60
11.7	20.9	9.2	78.63
16.0	23.8	7.8	48.75
6.2	15.5	9.3	150.00
70.7	111.2	40.5	479.77

The results in the first series, A, were obtained when the soil was plowed in the fall. Those in the second series, B, are the results obtained from fallowing the soil. The results in the third column give the gain of B over A, and in the fourth column this gain is expressed in per cent, obtained in the usual way by dividing the gain of B over A by the value of A and multiplying by 100. One method too commonly used for expressing the final results is to average such percentages. Summing the values in the fourth column and dividing by the number of tests we have an average of 79.96. Using this to evaluate the results, we would say that the results from the second method show a gain of 79.96 per cent over the first method. This is an incorrect interpretation, since percentages should not be averaged.

The data should be handled in another way, by obtaining the average of A and the average of B, or 11.78 and 18.53, respectively.

The gain of B over A is now 6.75, and dividing this gain by the average of A, 11.78, the result is 57.30 per cent. This is the correct percentage value to use when expressing the gain of B over A. It is to be noted that there is a big difference between the two percentage values. *In general, it is not a wise procedure to average percentages.*

This example illustrates the danger of averaging percentage values, and great care should be observed at all times in handling percentages. In this connection 'Student' has said that percentages should be used with the greatest care, they are fertile mothers of fallacy.

It is important, therefore, to use caution in the interpretation of results. This is especially true for those who may be using statistical analysis for the first time since, as already stated, there is a tendency for them to rely too much on the results of their analysis and too little on the common sense values. Statistical analysis and good judgment should go hand in hand, and we should never feel that the application of even the most refined methods of statistical analysis relieves us of using our own best judgment in the interpretation of the results. This does not mean that we would announce a certain conclusion if all of the evidence of the statistical analysis proved otherwise, but it does mean that we should not be carried away with the manifold methods of statistical analysis and rely entirely on figures alone. Sound judgment and statistical analysis should continue hand in hand until the final result is announced.

CHAPTER II

FREQUENCY DISTRIBUTIONS

As indicated in the previous chapter, when one is planning to make a statistical analysis of a particular lot of data it is necessary to obtain measurements or records of a sufficient number of individuals so that the data can be considered reliable. In other words, the population should be large enough to be truly representative of the condition it is planned to study. After one has measured or counted the characteristics of a large number of individuals, say several hundred, a thousand, or more, it is very difficult to form a definite idea of the nature of the material. One has many numbers recorded, but from studying these item by item he cannot draw any conclusion as to the population as a whole. It is clear, then, that it is desirable so to arrange the individuals as to obtain definite information regarding the population. The simplest way to do this is to arrange the various individuals in groups or classes. That is, in accordance with the nature of the material under observation one can decide what kind of grouping or classification he desires to follow and then record the individual items in the various groups according to the value or size of the character being studied. Such an arrangement of the individuals of a population in classes or groups is called a frequency distribution.

Number of Classes and Class Range or Interval. When a frequency distribution is to be made the first question to be decided is the number of classes to be used. This involves another point, the size of the classes, or the class range or class interval.

Regarding the number of classes, it is impossible to set a definite number that should be used in making frequency distributions. It would be arbitrary to fix an exact number, and furthermore, the nature of the different kinds of material that may be studied is so variable that it is impossible to establish a definite number of classes. For example, if one were counting the petals of a flower such as the common buttercup (*Ranunculus bulbosus*) it would be possible to have only a few classes since the number of petals on this plant usually vary from five to ten. Again, if the number of ray florets of a flower such as the common field daisy (*Chrysanthemum leucanthemum*) are being studied, they vary over a wide range. (See Table 11.) In such a study a large number of classes must be used, for if several classes were combined in a single class some of the important biological facts would be obscured. However, in dealing with data obtained by measuring or weighing, or even by counting such material as the number of seeds per plant, it is possible to group several measurements together and it is not necessary to have separate classes for each unit. Usually for most cases from 10 to 20 classes will be sufficient, but at times more may be needed. With certain kinds of material it is possible to have only a few classes. For example, when the number of culms per plant for wheat plants that have been grown under uniform conditions are grouped in classes, it is possible in certain cases to have only eight or ten classes. The same may be true with reference to the number of spikelets per head for wheat plants. Therefore, the nature of the material will quite largely determine the number of classes, and the important facts of the behavior of the material should not be obscured by having too few nor lost sight of by having too many classes. It may be stated with reference to the number of classes that the important consideration is to have a sufficient number to give a fairly uniform distribution.

The size of the class interval, then, will depend on the number of classes used. In deciding on the class interval the limits must also be considered carefully so that no confusion will arise in grouping the material in the various classes. While the limits are important, the same is also true of the mid-points of the classes, and it is well to consider both when deciding the class limits. That is, since after arranging the individuals in the various classes it is customary to assign to all the individuals in each class a value or weight equal to the mid-point of the class, the class limits should be so chosen as to leave no doubt as to what the mid-point is, and it should be readily determined. In making a frequency distribution of a population whose characters have been recorded by counting, for instance number of seeds, it is well to choose class limits whose mid-points are integers as, for example, 1–5, 6–10, 11–15, and the like. Here the mid-points will be 3, 8, 13, and so on.

When material that has been obtained by weighing or measuring is to be grouped, it is possible to use in some cases a plan similar to that suggested for numbers. Often, however, this is not so easily done and it is very important that the class limits be selected so that no confusion will arise as to the mid-points of the classes. This will be simplified if the matter is approached geometrically rather than arithmetically. Suppose one has a group size 1 to 5. If the class is arranged as follows:

1 2 *3* 4 5

it is clear that 3 is the mid-point since there are two spaces, from 1 to 2 and from 2 to 3, below it and the same number of spaces above it.

Suppose we have objects that have been measured in centimeters and we wish to arrange them in classes. If the data are of such a nature that some individuals are found with a very low value, say .1 or .3 of a centimeter, and other individuals are found measuring as high as 15 or 20 centimeters, we might arrange the first class to include all individuals whose measurements are more than 0.0 but less than 2.0 centimeters. The class limits then would be 0.0 as the lower limit and 1.9, or in reality 1.999+, as the upper limit. This means that all the individuals measuring up to but not equal to 2.0 centimeters would be placed in the first class. The mid-point of this class is then taken as 1.0 centimeter. For the second class the lower limit would be 2.0 and the upper limit 3.999+. This class would include all individuals measuring 2.0 or more but less than 4.0 centimeters, and the mid-point of this class is 3.0 centimeters. The other classes would be arranged similarly. In other words, in establishing class mid-points and class limits, the space included between the limits must be considered rather than the limits themselves. Thus, with the classes just considered, we have the following arrangement.

0.0 .5 *1.0* 1.5 2.0 2.5 *3.0* 3.5 4.0

As stated, the first class is from 0.0 to 1.999+, and it is clear that the mid-point is 1.0 since there are two spaces below and two spaces *above this mid-point. The mid-points of the other classes are determined similarly. This illustration further emphasizes the* statement above, that when determining mid-points we should consider the distance included within the class limits.

In stating the limits of the classes, then, they are usually handled as follows: 0.0–1.9, 2.0–3.9, 4.0–5.9, and so on, for the case just cited. Again, if we have material that has been weighed in grams and the classes differ by 3 grams, they would be arranged as follows: 0.0–2.9, 3.0–5.9, 6.0–8.9 grams, and so on.

It should be kept in mind that so far as possible the class interval should be the same for all classes. This will be illustrated very clearly in later chapters when we are dealing with the calculation of the various constants. Unless absolutely necessary the class interval should not be of varying sizes. It is possible in some

TABLE 1

YIELDS IN GRAMS OF 100 SOY BEAN PLOTS GROWN AT HSUCHOW, KIANGSU, IN 1933

230	264	274	236
291	276	309	241
290	327	330	235
273	335	304	216
292	318	216	217
297	344	300	200
256	369	257	286
246	317	241	249
258	265	235	310
312	192	223	306
448	239	229	288
328	342	242	250
274	310	218	268
309	211	252	229
295	186	274	350
375	214	262	312
336	225	340	329
312	275	302	393
279	305	233	271
246	269	283	290
239	216	292	318
288	237	261	318
310	269	306	240
234	152	223	283
316	305	217	297

cases that it will be necessary to change the class interval in certain classes in either the extreme lower or upper part of the frequency distribution, but if this is done it should be carefully noted.

Making a Frequency Distribution. The method of making a frequency distribution may be illustrated by the use of the data given in Table 1, page 14. These data are the yields in grams of 100 soy bean plots grown at Hsuchow, Kiangsu, in 1933. The first step is to decide on the number of classes and the class limits. The usual procedure would be to look through the data to determine the range of the measurements. From the data in Table 1 it is found that the lowest value is 152 grams while the highest is 448 grams. If we take the difference between these two measurements we find that the total range is 296 grams. By making use of this range we decide how many classes we will use and this will help determine the class interval, as follows. If we divide the total range, 296, by 30 we find that 30 is contained in 296 about ten times. This means that 10 or 11 classes would be sufficient to include all the individuals. Again, if we divide 296 by 25 we find that we will need 12 or 13 classes. If we decide to use this number of classes then the number that we have divided by may be taken as the class interval.

The next point to determine is the lower limit of the first class. As a general practice this will depend somewhat on the mid-point that may be desired. It is usually wise to select the lower limit of the first class so that the few individuals that may be in the first class would be fairly represented by the mid-point of that class. For the data in Table 1 we will select 140.0 as the lower limit of the first class and since we are using 25 as the class interval the upper limit will be 164.9+ grams. In practice it is not necessary to continue to write this value as 164.9+, as it is understood that all individuals weighing 140.0 grams but less than 165 grams are to be included in this class. The second class will then have the limits 165.0 to 189.9+, the third 190.0 to 214.9+, and so on until we have enough classes to include all the individuals. Since the largest weight is 448 grams we must continue our classes so that there is a class for this measurement, which means that we would continue our system of classification by intervals of 25 up to 440.0–464.9+. The complete number of classes with their class limits are given in Table 2.

When the classification system has been determined we would take each individual item in turn and record it in its proper class. For example, the first weight is 230 grams and it will be placed in class 215.0–239.9. The most convenient way is to make a mark

TABLE 2

FREQUENCY DISTRIBUTION FOR DATA IN TABLE I

CLASS VALUE V	NUMBER OF INDIVIDUALS IN EACH CLASS	TOTAL f
140.0-164.9	/	1
165.0-189.9	/	1
190.0-214.9	////	4
215.0-239.9	卌 卌 卌 卌	20
240.0-264.9	卌 卌 卌	15
265.0-289.9	卌 卌 卌 //	17
290.0-314.9	卌 卌 卌 卌 ///	23
315.0-339.9	卌 卌 /	11
340.0-364.9	////	4
365.0-389.9	//	2
390.0-414.9	/	1
415.0-439.9		0
440.0-464.9	/	1
		100

It should be noted that in arranging the classes for this frequency distribution the class limits are 140.0-164.9 grams, 165.0-189.9 grams, 190.0-214.9 grams, and so on, rather than 140-165 grams, 165-190 grams, 190-215 grams, and so on. The arrangement followed denotes clearly in which class an individual should be placed, while if the first class ended in 165 grams and the second class began with 165 grams it would not be clear in which class an individual weighing 165 grams would be placed.

for this individual. The second individual weight is 291 and it will be placed in class 290.0–314.9, and so on for all the individuals. If a mark is made for each individual that occurs in a particular class and if each four marks in the class are crossed with the fifth mark, it will be easy to sum the individuals in each class when the distribution is completed. Following this system, for these 100 individuals we have the results as given in Table 2, page 16.

In common practice it is customary to use the capital letter *V* to refer to the classes or class value, and in any computations this *V* refers to the mid-point of the class. The number of individuals in the different classes are designated by the letter *f*, or the frequency or number of times each class is represented. The total number in the population is designated by the letter *N*.

In making a frequency distribution it is very necessary that great care be taken that the marks are put in the proper classes. Such a system as just described does not permit of any method of checking except to make a second distribution. If the second distribution differs from the first there is no way to find what particular individual or individuals have been placed in the wrong class unless a third distribution is made, either for the total or for those particular classes where the discrepancies occur. For large numbers of individuals this system of recording the frequencies is also tedious. It is often true also that we have several measurements or characteristics for each individual of our population, and as we may want to study these characteristics in various ways it is often more convenient to copy all the data on cards. One card would be used for the measurements or characteristics for each individual in the population.

It may seem that the work of copying the records on cards will consume considerable time, but experience has shown that where one has three or four or more characters or measurements to be studied the time spent in copying the data on cards is more than saved as the data are studied later. There is also an added advantage in the card system, as the cards furnish a duplicate record of valuable data. The first record, which may have been taken on loose-leaf sheets or in a record book, can serve as a permanent record which should be kept in a safe place at all times. The data as recorded on cards will furnish the material for study. It is important, of course, that after being copied the data should be read back with the original for purposes of checking.

An illustration of how the card system is used is given here. Suppose we have the following data to be studied.

1.	Average height of plant, cm.	69.8
2.	Number of culms per plant	3
3.	Average number of spikelets per culm	34
4.	Total number of grains per plant	157
5.	Average number of grains per head	52
6.	Total weight of grains per plant, gms.	2.571
7.	Average yield of grains per head, gms.	.857
8.	Average weight of kernels in milligrams	16.376

For each of the individuals in the population we would have a similar record. One may have special cards printed with the headings to be used, but this is unnecessary and it often happens that it is desired to make changes in such headings from time to time. Plain cards may be used and instead of taking the time to write the headings for the characters to be studied it is sufficient to use numbers to designate the characters. It is even unnecessary to use numbers since the various characters may be recorded in order on the cards. If this latter system is followed then it is only necessary to have one guide card which would name in order the characters studied and would give the unit of measurement. The cards chosen for this work should not be too large, as they will not handle easily. A card three inches by five inches will take care of a good many characters. It might also be suggested that it is better to use a card than a light-weight paper, since papers of ordinary thickness do not shuffle or sort out so well as do cards of heavier paper. Cards of the thickness of ordinary library cards, or of even lighter weight, are very satisfactory. We may arrange the data on a card as follows:

1	*2*	*3*	*4*
69.8	3	34	157
5	*6*	*7*	*8*
52	2.571	.857	16.376

After recording the data on cards one would arrange his classes as before and then would sort out the cards, for one character at a time, placing them in piles in accordance with the system of classification and the value of the character for the particular grouping that is being made. It may be desirable to have small trays to hold the different cards, and to use slips of paper to indicate the class limits for each particular tray. It is not necessary to use these trays, as a little experience will enable one to do the sorting very conveniently without any elaborate system of trays.

When the sorting for the entire population has been completed the cards in each pile or group are counted and the number recorded in the proper class. The work may be checked very readily by counting back each pile of cards in turn. When checking in this way it is necessary to keep in mind only one set of class limits at a time. Thus, when checking for class 215.0–239.9 of our frequency distribution above it is necessary to keep only these numbers in mind, and if one card is found which belongs outside of this class it is a simple matter to place it in its proper class. In this way a frequency distribution can be made and checked much more quickly than by recording item by item as was done above. It may be desirable to keep the cards in the groups into which they have been sorted, in case they are needed for further study, and they may be tied together with twine or bound with rubber bands.

Types of Frequency Distributions. Frequency distributions are of various types, depending on the nature of the material under observation. One is the symmetrical distribution, such as is obtained in the case of tossing coins, where one side of the coin is designated heads and the other side tails, and recording the number of heads. Suppose eight coins are tossed and the number of heads turned up are recorded. It is possible to have various numbers of heads, from no heads, or all tails, to eight heads and no tails. If this is done a great many times a frequency distribution approaching the normal or symmetrical type will result. The result of tossing eight coins 2000 times and recording the number of heads is given in Table 3.

TABLE 3

DISTRIBUTION OF NUMBER OF HEADS OBTAINED BY TOSSING EIGHT COINS 2000 TIMES

NUMBER OF HEADS V	FREQUENCY OF OCCURRENCE OF NUMBER OF HEADS IN V f	RESULTS OBTAINED BY EXPANDING THE BINOMIAL 2000 $(\frac{1}{2}+\frac{1}{2})^8$
0	11	7.8125
1	62	62.5000
2	196	218.7500
3	421	437.5000
4	574	546.8750
5	467	437.5000
6	203	218.7500
7	55	62.5000
8	11	7.8125
	2000	2000.0000

This gives a frequency distribution approaching the normal or symmetrical type and is similar to the one that would be obtained by expanding the binomial 2000 $(\frac{1}{2}+\frac{1}{2})^8$. In other words, normal distributions may be obtained by expanding a binomial similar to the one given here.

A normal curve is also obtained from the data given by Yule, where the heights in inches have been obtained for 8585 men from England, Scotland, Wales, and Ireland. This distribution is given in Table 4.

TABLE 4

DISTRIBUTION OF HEIGHTS IN INCHES OF A NUMBER OF MEN FROM ENGLAND, SCOTLAND, WALES, AND IRELAND (YULE)

HEIGHT WITHOUT SHOES, INCHES	TOTAL FREQUENCY
57	2
58	4
59	14
60	41
61	83
62	169
63	394
64	669
65	990
66	1223
67	1329
68	1230
69	1063
70	646
71	392
72	202
73	79
74	32
75	16
76	5
77	2
	8585

Biological data do not always show symmetrical distributions, the asymmetrical or skew distributions being frequently obtained. A distribution of this type does not show a gradual rise until a middle value is found and then a similar decrease, but shows a greater tailing off on one side of the distribution than on the other. These distributions are very common and are of various kinds, ranging from those that are only slightly asymmetrical to those that have a class of greatest frequency and the observations sloping off to the one side of this class.

In Table 5, giving data on the heights in centimeters of 400 oat plants, arranged in classes differing by five centimeters, we have an asymmetrical type of distribution where the frequencies are not so uniformly grouped around the central values as they are in Tables 3 and 4.

TABLE 5

DISTRIBUTION OF HEIGHTS IN CENTIMETERS OF 400 OAT PLANTS

HEIGHT OF PLANT IN CENTIMETERS	FREQUENCY
45.0-49.9	2
50.0-54.9	9
55.0-59.9	20
60.0-64.9	35
65.0-69.9	91
70.0-74.9	125
75.0-79.9	91
80.0-84.9	26
85.0-89.9	0
90.0-94.9	1
	400

Another asymmetrical distribution is shown in Table 6. This is a distribution of 400 oat plants according to their weight of grain

TABLE 6

DISTRIBUTION OF WEIGHTS OF GRAIN IN GRAMS OF 400 OAT PLANTS

WEIGHT IN GRAMS	FREQUENCY
.00– .99	3
1.00–1.99	50
2.00–2.99	106
3.00–3.99	109
4.00–4.99	80
5.00–5.99	42
6.00–6.99	7
7.00–7.99	2
8.00–8.99	1
	400

TABLE 7

DISTRIBUTION OF WEIGHTS OF GRAIN IN GRAMS OF 500 OAT PLANTS

WEIGHT IN GRAMS	FREQUENCY
.00– 1.99	87
2.00– 3.99	192
4.00– 5.99	128
6.00– 7.99	71
8.00– 9.99	12
10.00–11.99	7
12.00–13.99	3
	500

in grams. The frequencies increase rapidly until the third class (2.00–2.99 grams) is reached, then the increase from the third to the fourth class is very slight. From the fourth class there is a very rapid decrease.

A more pronounced asymmetrical distribution is shown in Table 7, page 22. This is also a distribution of another lot of oat plants with respect to weight of grain. Here the rise is more rapid than in Table 6. In the distribution in Table 7 the greatest frequency is reached in the second class, then a gradual decrease is observed.

An example of another type of asymmetrical distribution is given in Table 8, showing the distribution of number of culms of 500 oat plants. The frequencies increase very rapidly to the fourth class, then decrease very gradually from the fifth and through the remaining classes.

TABLE 8

DISTRIBUTION OF NUMBER OF CULMS OF 500 OAT PLANTS

NUMBER OF CULMS	FREQUENCY
1	13
2	50
3	115
4	198
5	64
6	31
7	19
8	7
9	2
10	0
11	1
	500

TABLE 9

DISTRIBUTION OF NUMBER OF SPURS ON FLOWERS OF AQUILEGIA

NUMBER OF SPURS	FREQUENCY
5	449
6	187
7	66
8	12
9	4
10	1
	719

Data furnished by Dr. A. C. Fraser, Cornell University

An asymmetrical distribution of the extreme skew type is shown in Table 9, with the distribution of the number of spurs on the flowers of *Aquilegia canadensis*. In this distribution the first class has the greatest frequency and there is a gradual decrease to only one individual in class 10. This type of frequency is known as a *J*-type, since the curve resulting resembles the letter *J*.

There is another type of frequency distribution that is often extremely asymmetrical, known as the *U*-type. These are not very common but nevertheless do occur with certain kinds of data. An illustration, while not arising from a true frequency distribution, is given in Table 10, showing the per cent of cloudiness at Ithaca, New York. The data are the averages for six years. This distribution begins with a high percentage in January and gradually decreases until June. Then there is a gradual increase until another class of high percentage is reached in December.

TABLE 10

PER CENT OF CLOUDINESS AT ITHACA, NEW YORK. AVERAGE FOR SIX YEARS

MONTH	PER CENT OF CLOUDINESS
January	69.3
February	64.8
March	61.0
April	53.8
May	42.8
June	38.3
July	40.0
August	43.2
September	47.0
October	49.3
November	72.8
December	73.5

Data furnished by United States Department of Agriculture Weather Bureau, at Ithaca, New York

At times there are frequency distributions that are not only asymmetrical but are irregular in that they show an increase in frequencies up to a certain class, then a decrease followed by another increase until a second high peak is reached. Often several irregularities of this sort occur, depending on the nature of the data. Distributions of this kind are found in the variation of the ray florets of the common field daisy, as illustrated in Table 11 with the counts of the number of ray florets on the heads of daisies from the same plot of ground on different dates in the same year.

TABLE 11

DISTRIBUTION OF THE RAY FLORETS OF THE COMMON DAISY FROM THE SAME PLOT OF GROUND ON DIFFERENT DAYS

NUMBER OF RAY FLORETS	FREQUENCIES ON JUNE 23	FREQUENCIES ON JUNE 27	FREQUENCIES ON JULY 1
4			2
5		1	3
6	1	3	9
7	0	4	5
8	2	12	12
9	6	13	20
10	4	17	25
11	7	50	47
12	18	70	67
13	33	126	132
14	22	97	75
15	28	100	68
16	34	109	85
17	56	108	93
18	52	114	75
19	66	153	70
20	116	170	118
21	185	228	113
22	89	97	45
23	40	52	21
24	40	40	18
25	35	26	13
23	19	20	14
27	20	13	6
28	13	18	4
29	15	16	4
30	17	7	5
31	13	2	4
32	5	1	1
33	8	3	2
34	3	3	
35	1		
TOTAL	948	1673	1156

There is a decided grouping around the number 21 and a secondary grouping around number 13. In such irregular distributions, as pointed out earlier, it is unwise to change the classification system by combining classes since one might obscure important biological facts. If the classes were doubled in the case of the data in Table 11 one could not be sure of the point of greatest frequency.

Therefore, in cases of this sort it is important to keep a large number of classes.

The foregoing distributions illustrate some of the various types that are found with different kinds of data. The examples show gradations from the normal type of distribution to the extreme *J*-types and *U*-types. One may expect, therefore, to find various kinds of frequency distributions between these extremes depending on the nature of the data that are being analyzed.

CHAPTER III

GRAPHIC ILLUSTRATION

Graphic illustration is important in presenting statistics in such a way that they will convey to the eye quickly and clearly the facts or tendencies of the data under observation. It is very important to be able to present statistical information in some graphic form that will not only present the facts but will show them in an interesting way. Much has been written and many suggestions made as to methods of illustrating pictorially the facts and results of statistical studies. It is not the purpose in this short chapter to give any more than a few suggestions as to methods that may be used with the type of analysis discussed in these chapters. The student who desires to go further is referred to more extensive publications on the subject.

Of course, if one has material of a simple nature such as the height of men or length of wheat heads, the easiest way is to present a picture or photograph of the material. Often this is not possible, and one may obtain the same effect by a line diagram, in which, in accordance with a proper scale of measurement, each individual object is represented by the length or height of a line. Such an illustration is given in Figure 1.

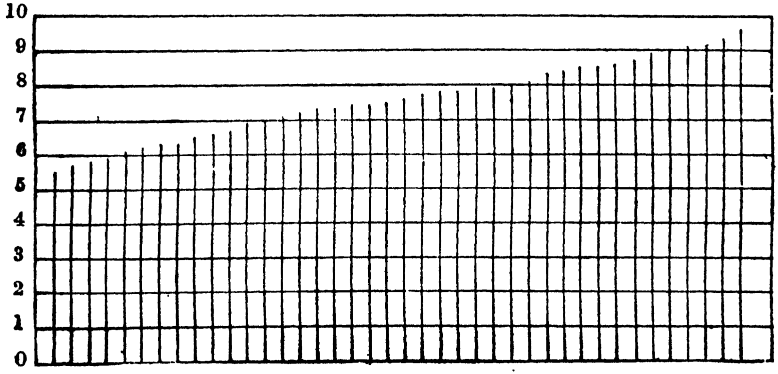


FIG. 1. Diagram illustrating the length of 40 wheat heads.

The data represented here are the measurements of the length of 40 wheat heads. The measurements were taken in centimeters, and the length *of the line represents the actual length of the head* measured. *In this arrangement the first line represents the shortest head, the second line the next longer, and so on until the last* line represents the longest head. It would also be possible to represent the length of head by a dot and connect the dots by a line. This method of graphic illustration has been termed an *ogive*. When the population is small this method of illustration is satisfactory, but if one had several hundred wheat heads, for instance, it becomes rather laborious and does not convey any clearer picture than other methods of illustration.

Another method of graphic illustration, which is similar to the one just described where the length of line represents a particular item, is that known as the bar diagram. In this method the value of each item is represented by the length of a bar. The advantage of the bar diagram over the line diagram is that the bar diagram tends to emphasize the facts since the width of the bar is wider than a mere line. The use of the bar diagram is illustrated in Figure 2.

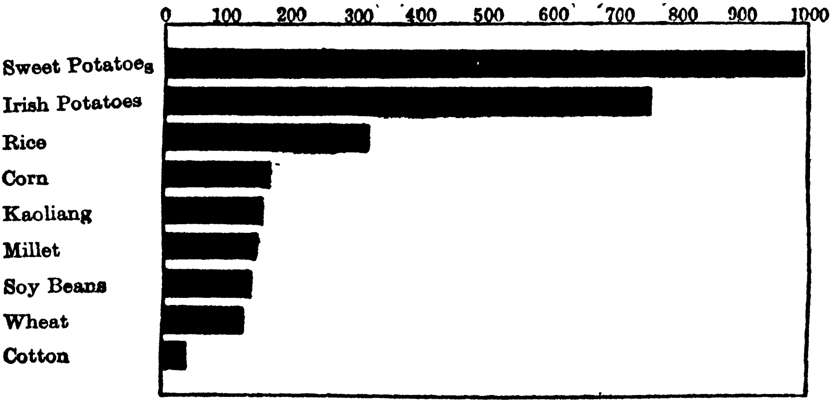


FIG. 2. Diagram showing the production in catties per mow of some of the important crops in China. Data from C. C. Chang, in *The Statistical Monthly*, January–February, 1932.

The data presented in this diagram are the yields in catties per mow of several of the important crops in China. The length of the bar for the various crops illustrates graphically the yields obtained from these different crops. As stated, one value of the bar diagram is that it tends to emphasize the facts that it is desired to illustrate. Another advantage is that it is possible to have each bar represent more than one type of information, as, for example, it would be possible to extend this chart to give the yield of rice straw and rice grain, or of seed cotton and lint cotton, and so on.

When one is dealing with data that have been arranged in frequency distributions, as discussed in Chapter II, it is often more convenient to illustrate the data by what is known as a frequency curve. The frequency curve will often convey the facts of the frequency distribution, or the tendency for the individuals to vary, better than mere observation of the numbers themselves. This is particularly true for the person who is not accustomed to thinking in terms of figures, and a line diagram will give him a much better idea of the data than the figures given in a frequency distribution.

The ordinary way of illustrating the data from a frequency distribution is to plot on some convenient coördinate paper the values corresponding to the classes and the frequency of the classes on what we may call the x and y axes. The classes, either class limits or class centers, are marked off on the horizontal or x axis and are called the abscissal distances, and the frequencies of the classes are marked off on lines drawn perpendicular to the x axis by using some convenient scale of measurement. These perpendicular lines are referred to as ordinates and the frequency of a particular class, represented by the area between the class limits, is sometimes referred to as an ordinate value. As an illustration of the plotting of a frequency curve the data in Table 6 in Chapter II have been used for plotting the curve in Figure 3, page 30.

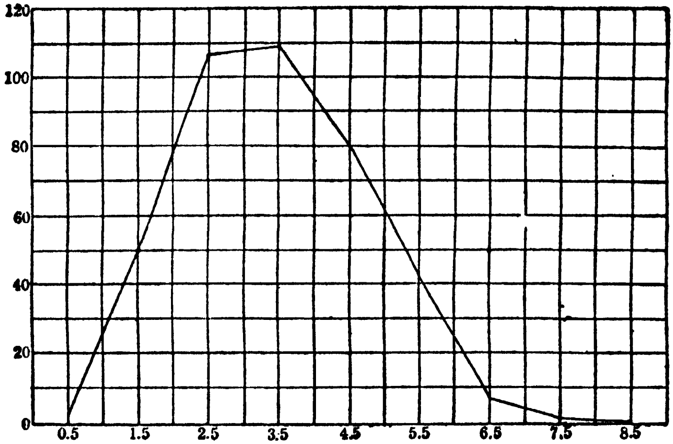


Fig. 3. Graphic illustration of the weights of grain for 400 oat plants. Data from Table 6 in Chapter II.

The class values are represented in order on the base line by the mid-points of the classes, beginning with the first class, and then on the ordinates erected on the mid-points of the several classes distances are located indicating the frequency of the classes. In locating these points any convenient scale may be used, such as letting each centimeter in height represent 10 individuals, or any convenient system may be used that makes it easy to locate the various points on the several ordinates. With the scale used, say for the first class of Table 6, a point is marked off indicating the three individuals of this particular class. For the second class, which centers at 1.5, the number of individuals is 50, and this is located on the proper ordinate, and so on for all of the different points. These points are then connected by straight lines, and when completed this gives what is known as a frequency curve. In this case it is often thought of as a broken-line curve, that is, the data are illustrated just as they occur without any attempt to smooth the curve. This method of plotting a curve is sometimes referred to as the method of *loaded* or *weighted* ordinates.

Another method of illustrating a frequency distribution is sometimes referred to as a histogram, or illustrating by what is known as the system of rectangles. This method is shown in Figure 4, with the same data as used in Figure 3.

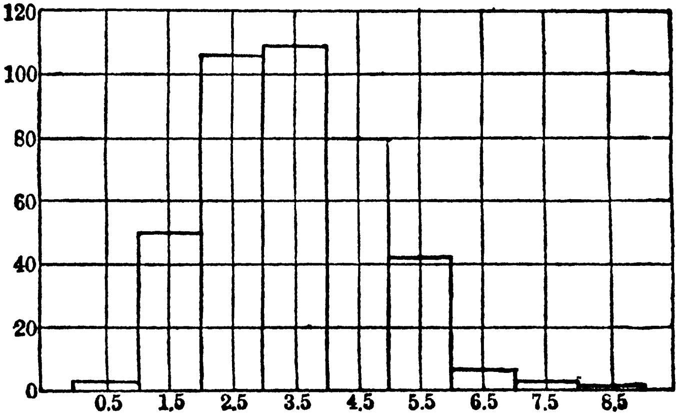


FIG. 4. Histogram for the same data as illustrated in Figure 3.

This method of plotting a frequency curve consists of locating the distances on the proper ordinates, as before, but instead of connecting the several points by means of a straight line a rectangle is erected on the ordinate by first drawing through the point as located a line horizontal to the base line and extending half way to the next class center, or half way to the ordinate representing the mid-point of the next class. Then lines from the end of this line are drawn perpendicular to the base line, thus completing the rectangle. With this method of illustration rectangles are erected on all of the ordinates in accordance with the location of the points on the basis of the scale of measurement.

It is purely a matter of choice as to which method is used to illustrate a frequency distribution. It may be possible that in some cases the system of rectangles will emphasize the frequencies at the extremes of the distribution a little more than does the frequency curve. If the system of rectangles is to be used it is better to follow the method shown in Figure 5, which is a graphic illustration of the same data as used in Figures 3 and 4.

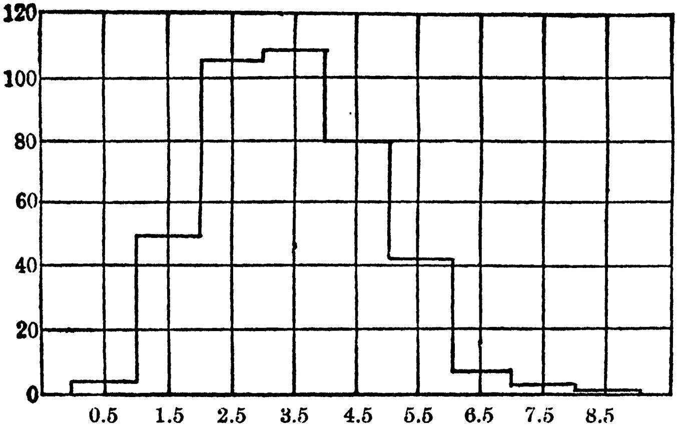


FIG. 5. Histogram with inner lines eliminated for the same data as illustrated in Figure 4.

The points for drawing the rectangles are located as in Figure 4, but instead of completing the rectangles the inner lines are omitted. This will be clear on referring to Figure 5. This method is less confusing than that in Figure 4, since there are fewer lines and therefore the picture is more clearly brought out.

It may often be important to illustrate more than one frequency distribution on the same diagram, and it may be possible that the size of the population of the different frequency distributions or the number of observations is not the same in each case. When this is true and the curves are plotted as explained above by marking off on the ordinates points indicating the frequency of each class, that is by using the actual numbers in the different classes, one would have a diagram similar to that shown in Figure 6, page 33.

In this illustration, due to the difference in size of the populations, we do not have as accurate an idea of the relation between the two distributions as we would if the effect of the difference in the size of the populations were eliminated. This may be done by calculating the frequency in each class as a percentage of the whole population, and then plotting the percentages for each class rather than the actual numbers. This has been done in Figure 7, using the same data as in Figure 6.

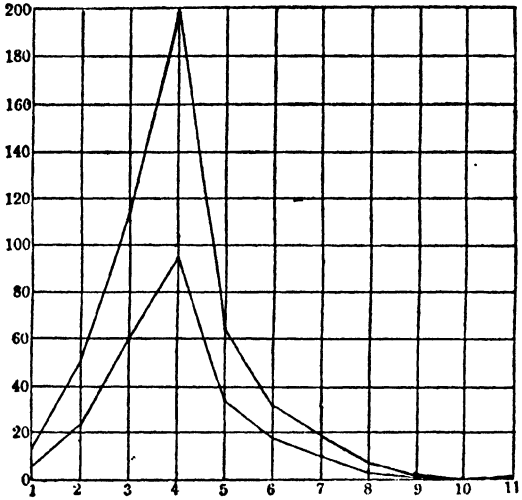


FIG. 6. Graphic illustration of the data from Table 8 in Chapter II, which are the number of culms for 500 oat plants. The lower curve is for one-half of the population and the upper curve is for the total population.

On observing Figure 7 it is clear that a better idea of the two distributions is obtained than is possible from Figure 6. In plotting the two graphs in Figure 7 the solid line is used for one and the dotted line for the other. If one has more than two curves to include in the same figure a dot-and-dash line may be used for the third distribution, and perhaps a long dash line if one chooses to illustrate four distributions in one figure.

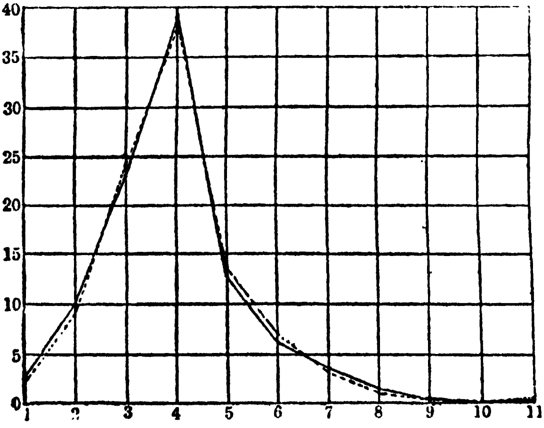


FIG. 7. Illustrating the use of percentages in plotting the same data as shown in Figure 6.

Another illustration of plotting a frequency curve is given in Figure 8, page 34. The data used are from Table 3, in Chapter II.

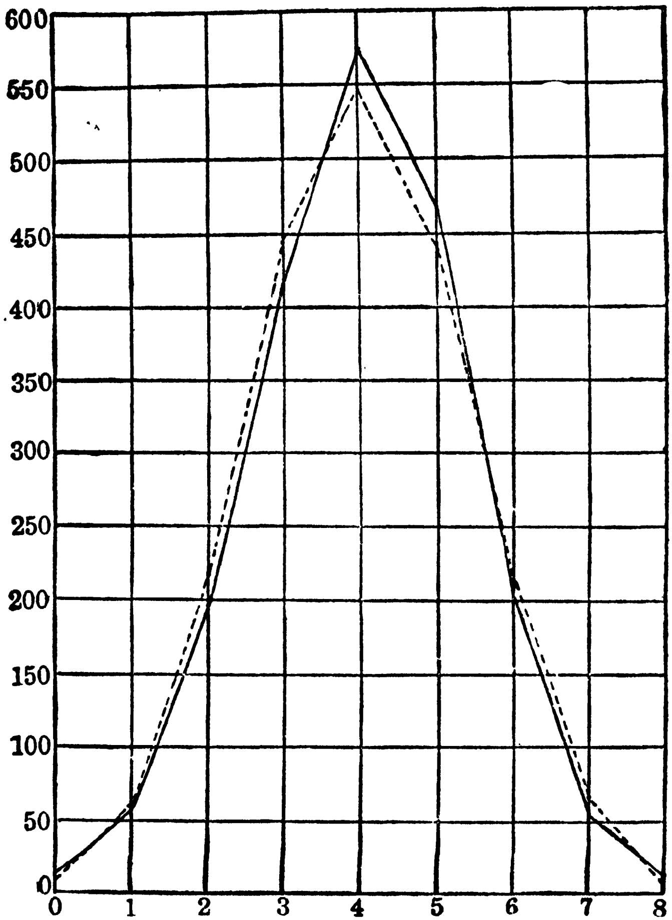


FIG. 8. Graphic illustration of the data from Table 3 in Chapter II, which are the results of tossing eight coins 2000 times. The solid line represents the results obtained, and the dotted line represents the results expected by expanding the binomial 2000 $(\frac{1}{2}+\frac{1}{2})^8$.

Two curves are included in this diagram. The solid line represents the actual results and the dotted line represents the results expected if the observed results had agreed with those obtained from expanding the binomial 2000 $(\frac{1}{2}+\frac{1}{2})^8$. The two curves agree rather closely, and show that the results from experience, or experiment, agree rather well with those expected. The difference, as shall be discussed later, is due to chance variation.

An illustration of a frequency curve for a decidedly skew distribution is given in Figure 9.

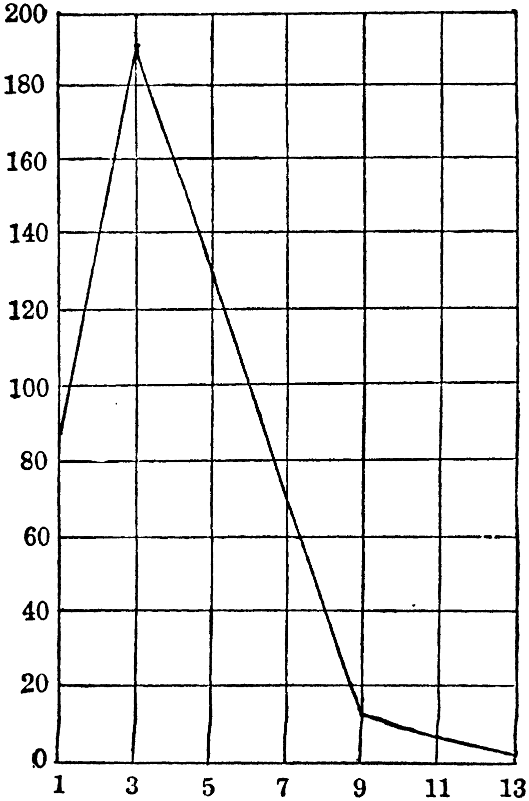


FIG. 9. Graphic illustration of the distribution of weights of grain for 500 oat plants. Data from Table 7 in Chapter II.

The data illustrated are taken from Table 7 in Chapter II, and are the weights of grain for 500 oat plants. It is seen that the curve starts with a rather high frequency in the first class and reaches a maximum frequency in the second class, and from this point there is a gradual decrease.

A case of an extremely skew, or asymmetrical, curve is illustrated in Figure 10.

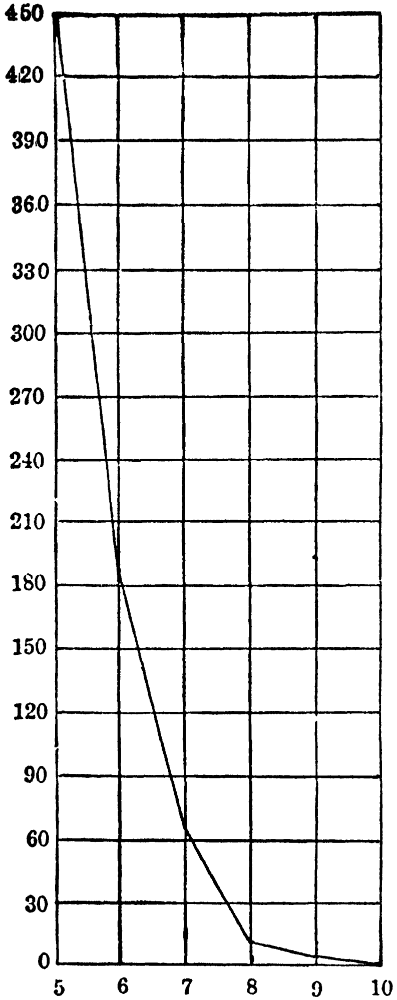


FIG. 10. Illustrating the distribution of number of spurs on *Aquilegia* flowers. Data from Table 9 in Chapter II.

The data are from Table 9 in Chapter II. In this distribution the curve starts at the first class, which is the class of greatest frequency, and decreases rapidly until the last class is reached. On account of the general shape of this frequency curve, which resembles the letter *J* of the English alphabet, it is referred to as a *J*-shape curve.

There will be additional methods used for graphic illustration in the course of the discussions following, but they will be similar to the ones discussed here. As already stated, there are other ways of making graphic illustration, but for the type of problems with which we are most concerned the more important have been shown here.

For the plotting of data, by either the frequency curve or histogram, there are certain considerations or rules that should be observed, as follows:

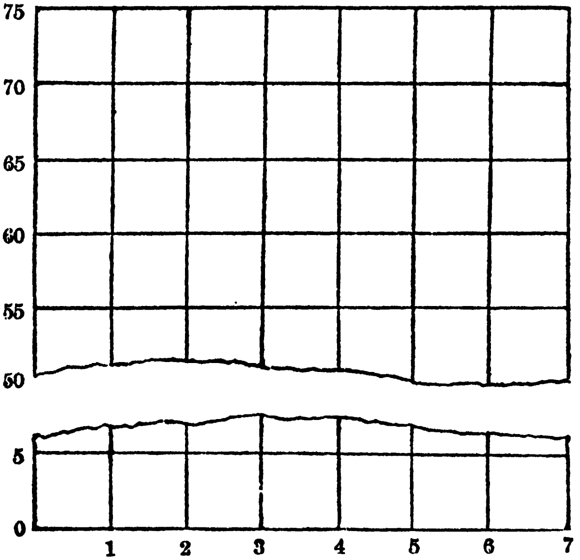
The choice of scale is important as the completed curve should be drawn to such a scale that it does not exaggerate nor minimize the facts it is intended to show. That is, the height of a figure should not be too great in proportion to its width, nor should the width be made so great as to minimize the differences between the several classes, which may be the important fact that it is intended to illustrate.

When several diagrams are used to illustrate similar results it is best, so far as possible, to use the same scale of measurement in the plotting, so that the curves may be directly compared with each other.

It is important not to put too many curves on one diagram, usually not more than four or five, especially if the curves overlap considerably. If a number of curves are included on one diagram then the entire diagram loses interest, since it forces the reader to trace carefully the different lines in order to follow the various frequency distributions. Therefore a number of graphs on the same figure should be avoided.

When plotting curves including populations of varying numbers of observations it is better to determine the percentages and plot the curves on the basis of the percentages for the different classes.

So far as possible, the zero line should always be shown. There may be occasions when it will require too much space to start the base of the diagram at zero, and in these cases it would be well to indicate the zero line by a break in the diagram, as illustrated in the following example. When this



Illustrating break in diagram to indicate zero line.

> is not done there are times when we may over-emphasize the tendency of the curve to increase in a certain direction.
>
> Unless for some very special reason, it is better for the horizontal scale to read from left to right and for the vertical scale to read from the bottom to the top. In preparing a diagram it is better to give the figures for the abscissal values at the bottom of the diagram and the grouping for the ordinate values to the left of the diagram.
>
> One should not, in order to economize on space, plot frequency curves on such a small scale that they lose the effect they are intended to convey. All the figures and lettering must be carefully done so as to convey a clear idea of what the diagram is intended to show.

The foregoing discussion and illustrations, together with some of the general rules to be followed, give some idea of the methods used for ordinary graphic illustration. For more elaborate illustrations various methods have been devised which are particularly useful. If the reader is interested in further information he may consult the more extensive discussions on this subject.

CHAPTER IV

CONSTANTS OF POSITION

In Chapter II we have discussed the method of classifying individuals into groups and have shown that the type of frequency distribution so obtained may vary with the nature of the character of the material. While such frequency distributions are useful and furnish some information regarding the data being studied, what is needed is more than mere classification. A frequency distribution gives some idea of the range of the material, but it is not possible from a study of a frequency distribution to make any definite quantitative statements regarding the nature of the material as a whole. We might have two frequency distributions of similar material having the same range, and yet they may differ with respect to the tendency of the individuals to be grouped around certain central values. For example, the two distributions given below will illustrate this point. The data are yields per culm for two different varieties of oats.

Distribution of Yields Per Culm of Oats in Decigrams

Class	Variety A	Variety B
0.0–0.9	78	153
1.0–1.9	96	83
2.0–2.9	72	40
3.0–3.9	37	15
4.0–4.9	12	5
5.0–5.9	4	1
Total	299	300

It is evident that in comparing these two frequency distributions we need to know more than the range of the material. It is not enough to state that for Variety A there is a different arrangement of the frequencies than for Variety B.

What is needed, therefore, in addition to knowing the range, is to know something concerning the tendency of the individuals to center about certain values, or the position of the individuals, and some expression that will give a measure of this position. Such measures of position are called averages, and they define the position of the distribution in the same terms or values that have been used in the study of the variable. For example, if we are studying the length of a variable which has been measured in centimeters, then the average is expressed in centimeters, or if the weight of a variable has been studied and determined in grams the average is expressed in grams, and so on for other units of measurement.

Yule mentions some of the desirable points regarding these position constants, or averages, as follows: "(*a*) In the first place, it almost goes without saying that an average should be rigidly defined, and not left to the mere estimation of the observer. An average that was merely estimated would depend too largely on the observer as well as the data. (*b*) An average should be based on all the observations made. If not, it is not really a characteristic of the whole distribution. (*c*) It is desirable that the average should possess some simple and obvious properties to render its general nature readily comprehensible: an average should not be of too abstract a mathematical character. (*d*) It is, of course, desirable that an average should be calculated with reasonable ease and rapidity. Other things being equal, the easier calculated is the better of two forms of average. At the same time too great weight must not be attached to mere ease of calculation, to the neglect of other factors. . . ." It is also desirable that the average should not be affected seriously by methods of grouping or fluctuations of sampling, and finally it is very important that the average may be treated algebraically.

The most common constants for determining position are the *arithmetic mean*, the *median*, and the *mode*. To these may be added the *geometric mean*, which has certain special uses. These constants may be defined as follows.

The *arithmetic mean* is the value obtained by summing all of the individual items and dividing by the total number of items.

The *median* of a series may be expressed as the value of that individual in a population which divides the distribution into two equal parts.

The *mode* is the value that occurs most frequently in the distribution, or the most common measurement. There are two types of mode which may be considered, the empirical mode and the calculated or theoretical mode. These will be discussed in detail later.

The *geometric mean* of a frequency distribution is the result obtained by multiplying together the values of all the items and determining the *nth* root of their product.

These constants of position will now be discussed in the order named.

Arithmetic Mean. As indicated by the definition, the arithmetic mean, or more briefly the mean, is obtained by summing the values of all the individuals and dividing this sum by the total number of individuals. It is easily determined and readily understood, and it may be noted that it fulfills the first two conditions suggested by Yule. It is evident that the arithmetic mean is not dependent upon any particular method of grouping but depends chiefly on the values of the items concerned. It should be pointed out that one of the characteristics of the mean is that often its value does not agree with any particular individual of a group. For example, suppose that 100 men students were measured as to height. The value for the mean may be 67.56 inches, while it is entirely possible that not one of the 100 measured would be exactly this height.

When the number of individuals being studied is small the most direct way to obtain the mean is to sum the values of all the individuals and divide by the total number. For large numbers of individuals it is more convenient to determine the mean from a frequency distribution. In calculating the mean from a frequency distribution the several individuals in any one particular class are considered as having a value equal to the mid-point of that class. For example, in Table 12 the two individuals in the first class are assigned the value of 137.5 grams, since this is the mid-point of that class. The individuals in the remaining classes are treated similarly. That such a process under ordinary circumstances does not introduce an appreciable error is made clear by the following example with yields of soy beans from 100 plots. The mean determined by summing the individual plot yields and dividing by the total number of plots is 276.460 grams, while the mean obtained by grouping the 100 yields into classes differing by 25 grams is 276.500 grams. The difference between the two means is only .040 grams.

In some cases there may be differences somewhat larger than this one, but in general with the usual types of frequency distributions the difference between what may be called the true arithmetic mean (obtained by summing the values of all the individuals and dividing by the total number) and the mean obtained by grouping will be slight. This is on the assumption that when the population studied is sufficiently large the individuals in each class will be centered about the mid-point of the class. There are exceptions to this condition if one should have the types of curves where the measurements give the shapes known as the *J*-type or the *U*-type curves, because under such conditions there would be a tendency for the individuals to pile up more in that part of the class toward the class of greatest frequency. In such cases the individuals would not be equally distributed around the mid-point. Such types of distribution are rather unusual, and for those that are symmetrical or only moderately asymmetrical the individuals of a class will be found to be rather evenly distributed around the mid-point, except in the extreme classes where only a few individuals are found. For certain distributions there may be a tendency for the individuals in the classes below the class of greatest frequency to pile up somewhat in the upper half of the class, but any effect that might be caused by this condition would be smoothed out by the opposite tendency for the individuals in the groups above the class of greatest frequency. In the final result the mean for an average distribution where a sufficient number of individuals have been studied will agree closely with the value obtained by summing all the items and dividing by the total number.

The method of calculating the mean from a frequency distribution is shown in Table 12.

TABLE 12

METHOD OF CALCULATING THE MEAN FROM A GROUPED FREQUENCY DISTRIBUTION. DATA ARE YIELDS IN GRAMS OF SOY BEAN PLOTS GROWN AT HSUCHOW IN 1933

CLASS	V	f	fV
125.0–149.9	137.5	2	275.0
150.0–174.9	162.5	2	325.0
175.0–199.9	187.5	15	2812.5
200.0–224.9	212.5	42	8925.0
225.0–249.9	237.5	69	16387.5
250.0–274.9	262.5	78	20475.0
275.0–299.9	287.5	67	19262.5
300.0–324.9	312.5	57	17812.5
325.0–349.9	337.5	42	14175.0
350.0–374.9	362.5	18	6525.0
375.0–399.9	387.5	5	1937.5
400.0–424.9	412.5	2	825.0
425.0–449.9	437.5	1	437.5
		$N=400$	$\Sigma fV=110175.0$

$$M=\frac{110175.0}{400}=275.4375 \text{ grams}$$

The method illustrated in Table 12 is simple in that the mid-point of a class is taken to represent the value of all individuals in that class. This mid-point or value (V) is multiplied by the number of individuals in the class, giving fV. The values in column fV are summed and the total divided by 400, giving the mean equal to 275.4375 grams. The formula for the mean (M) calculated by this method is

$$M=\frac{\Sigma fV}{N}$$

This process sometimes requires the multiplication of large numbers and hence is laborious, but a shorter method may be used. The details of this method are given in Table 13.

TABLE 13

METHOD OF CALCULATING THE MEAN BY THE SHORT METHOD

Assumed Mean $(G) = 262.5$

CLASS	V	f	$(V\text{-}G)$ or D	fD
125.0–149.9	137.5	2	−125	− 250
150.0–174.9	162.5	2	−100	− 200
175.0–199.9	187.5	15	− 75	−1125
200.0–224.9	212.5	42	− 50	−2100
225.0–249.9	237.5	69	− 25	−1725
250.0–274.9	262.5	78	0	
275.0–299.9	287.5	67	25	1675
300.0–324.9	312.5	57	50	2850
325.0–349.9	337.5	42	75	3150
350.0–374.9	362.5	18	100	1800
375.0–399.9	387.5	5	125	625
400.0–424.9	412.5	2	150	300
425.0–449.9	437.5	1	175	175
		$N=400$		10575
				−5400
				$\Sigma fD=5175$

$$c=\frac{5175}{400}=12.9375$$

$$M=262.5+12.9375=275.4375 \text{ grams}$$

With this method we assume some value as the mean, as, for example, the mid-point of class 250.0–274.9, or 262.5. This value is referred to as the *guess* at the mean, and is designated by G. The deviation of the mid-point of each class from this guess is obtained and recorded in column D. It must be kept in mind that the deviations of the classes below the guess, or to the left of the assumed mean if the curve is plotted, are recorded as minus, and those above, or to the right of the assumed mean, as plus.

The next step is to multiply each value in column D by its corresponding frequency and record the result in column fD with the proper sign. The algebraic sum of the values in fD is then obtained and this sum, divided by the number of individuals, gives a correction value, c, for the mean. This correction value shows how far the guess, or assumed mean, differs from the true mean. If c is *plus* it means that we have guessed too low, and the value of c must be added to the guess to obtain the true mean. If c is *minus* it indicates that the guess was too high, and the value of c must then be subtracted to obtain the true mean. That is, in this particular case, the mean was assumed to be at 262.5 grams, but the calculation shows this to be too low by 12.9375 grams. In other words, we guessed too low, and the value of this correction must be added to the assumed mean to obtain the true mean, 275.4375 grams.

This method is preferable to the one shown in Table 12, since we are dealing with smaller numbers in the product column (fD) and the multiplication and addition are simpler. It should be kept in mind that as a rule the smaller the numbers to be used in calculations the greater will be the accuracy of the results. For example, there is more chance of making an error in multiplying 137.5 by 2 than there is in multiplying 125 by 2. For each additional figure to be multiplied there is a chance for error.

The formula for the mean as calculated by the second method is

Let $G=$ assumed mean

$c=$ correction, or $\frac{\Sigma fD}{N}$

The sign of c may be either *plus* or *minus*.

We can write the formula

$M=G+c$, thus obtaining the algebraic sum.

A more convenient method and the one commonly used in calculating the mean is one in which the class intervals are considered as unity. This is called the unity-step method. It is illustrated by Table 14, page 46, with the same data as were used in Tables 12 and 13. This method is the one recommended for use in preference to all others, as it will prove to be less laborious and since smaller numbers are used in the calculations there are fewer chances for making errors.

This method is much shorter than the one illustrated in Table 13 and differs from it in that instead of measuring the deviations of the classes from the assumed mean by actual differences they are measured on the basis of unity values. In other words, since in most frequency distributions the class interval is the same for each class, or the classes are separated by an equal amount, we assume the simplest difference to measure these class intervals.

TABLE 14

METHOD OF CALCULATING THE MEAN BY THE UNITY-STEP METHOD

Assumed Mean $(G) = 262.5$

CLASS	f	$(V-G)$ D	fD
125.0–149.9	2	−5	−10
150.0–174.9	2	−4	− 8
175.0–199.9	15	−3	−45
200.0–224.9	42	−2	−84
225.0–249.9	69	−1	−69
250.0–274.9	78	0	
275.0–299.9	67	1	67
300.0–324.9	57	2	114
325.0–349.9	42	3	126
350.0–374.9	18	4	72
375.0–399.9	5	5	25
400.0–424.9	2	6	12
425.0–449.9	1	7	7
	$N = 400$		423
			−216
			$\Sigma fD = 207$

$c = \Sigma fD/N \times ci \qquad c = \frac{207}{400} = .5175 \times 25 = 12.9375$

$M = G + c \qquad M = 262.5 + 12.9375 = 275.4375$ grams

Since the simplest difference is one, then each class is assumed to be one unit or one step removed from the preceding class. For this reason it is important that frequency distributions be so arranged that the class interval is uniform throughout the distribution.

In the unity-step method we again guess at the mean, and assume each class in turn to differ from this guess, G, by one unit. These differences are recorded in column D and are then multiplied by the corresponding frequency, having regard to the proper sign, giving the column fD. The values in fD are summed algebraically and the result is divided by the total number of individuals, giving a correction value, c. This value c is multiplied by the class interval and then added to the assumed mean to obtain the true mean. We obtain the same value for the mean (275.4375) as by the two preceding methods, and the calculations are much simpler. For instance, by the unity-step method the values in column *D* are -5, -4,-3, and so on, while by the method in Table 13 they are -125, -100, -75, and so on, and it is easier to obtain the values for column *fD* from the smaller deviations.

The formula for obtaining the mean by the unity-step method is to obtain *c* as before.

Then

$$M=G+(c\times \text{class interval, or } ci)$$

In Table 15, using the same data as in Table 14, the guess has been taken at 337.5 and the same value for the mean is obtained. It is noted that in this case the sign of *c* is *minus*, therefore the numerical value of *c* must be subtracted from *G* to obtain the mean. This merely shows that it makes no difference what value is taken as the assumed mean so long as the proper signs and correction are

TABLE 15

METHOD OF CALCULATING THE MEAN BY THE UNITY-STEP METHOD
Assumed Mean $(G)=337.5$

CLASS	*f*	$\frac{(V\text{-}G)}{D}$	*fD*
125.0–149.9	2	-8	- 16
150.0–174.9	2	-7	- 14
175.0–199.9	15	-6	- 90
200.0–224.9	42	-5	- 210
225.0–249.9	69	-4	- 276
250.0–274.9	78	-3	- 234
275.0–299.9	67	-2	- 134
300.0–324.9	57	-1	- 57
325.0–349.9	42	0	
350.0–374.9	18	1	18
375.0–399.9	5	2	10
400.0–424.9	2	3	6
425.0–449.9	1	4	4
	$N=400$		38
			-1031
			$\Sigma fD=-993$

$$c=\frac{-993}{400}=-2.4825\times 25=-62.0625$$

$$M=337.5+(-)\ 62.0625=275.4375$$

regarded. Assuming the mean in different classes also affords a method for checking the calculation of the mean. Often computers may find it desirable to assume the mean in the first class, especially if the frequency distribution has only a few classes, and in this way the use of minus signs is avoided.

Another example to show the unity-step method for determining the mean is given in Table 16. The data are the weights in grams of 400 wheat plots grown at Hsuchow, China, during the season of 1932–33.

TABLE 16

METHOD OF CALCULATING THE MEAN BY THE UNITY-STEP METHOD. DATA ARE WEIGHTS IN GRAMS OF WHEAT GROWN AT HSUCHOW IN 1932–33

Assumed Mean $(G) = 412.5$

CLASS	f	$(V\text{-}G)$ D	fD
250.0–274.9	4	−6	−24
275.0–299.9	3	−5	−15
300.0–324.9	13	−4	−52
325.0–349.9	27	−3	−81
350.0–374.9	42	−2	−84
375.0–399.9	55	−1	−55
400.0–424.9	66	0	
425.0–449.9	59	1	59
450.0–474.9	43	2	86
475.0–499.9	44	3	132
500.0–524.9	17	4	68
525.0–549.9	13	5	65
550.0–574.9	7	6	42
575.0–599.9	6	7	42
600.0–624.9	1	8	8
	$N = 400$		502
			−311
			$\Sigma fD = 191$

$$e = \frac{191}{400} = .4775 \times 25 = 11.9375$$

$$M = 412.5 + 11.9375 = 424.4375$$

The result of the calculation is

$$c = .4775 \times 25 = 11.9375$$
$$M = 412.5 + 11.9375 = 424.4375$$

Thus the mean for this distribution is 424.4375 grams.

In order to show the effect of grouping, or of arranging the classes differently, on the mean, we may use the same data as given in Table 12. The same individual plot yields have been grouped by three other systems of classification, namely by class intervals of 30, 40, and 50 grams. These distributions are shown in Table 17, page 50, and the same table is used later in the discussion of the mode.

The means obtained are as follows:

CLASS INTERVAL 25 GRAMS	CLASS INTERVAL 30 GRAMS	CLASS INTERVAL 40 GRAMS	CLASS INTERVAL 50 GRAMS
Mean = 275.4375	276.275	276.400	275.500

It is seen that although the class interval has been changed considerably and even in one case has been doubled, as compared with the original distribution in Table 12, still the mean is only slightly affected.

A very important advantage of the mean as a constant of position is that it may be treated algebraically. That is, when we have the means of several series of similar material (M_1, M_2, etc.) and the number of the population in each case (n_1, n_2, etc.) we can determine the mean of the whole series by the following formula

$$NM = n_1M_1 + n_2M_2 + n_3M_3 \text{ etc., and}$$
$$M = \frac{n_1M_1 + n_2M_2 + n_3M_3 \text{ etc.}}{N}$$

For example, with the data used in Table 12 the population was divided and the means of the first 200 plots and the last 200 plots were obtained. The mean of the first 200 plots is 273.625 and for

TABLE 17

EFFECT ON THE MEAN BY DIFFERENT METHODS OF GROUPING

CLASS INTERVAL 25 GRAMS	f	CLASS INTERVAL 30 GRAMS	f	CLASS INTERVAL 40 GRAMS	f	CLASS INTERVAL 50 GRAMS	f
125.0–149.9	2	125.0–154.9	2	125.0–164.9	3	125.0–174.9	4
150.0–174.9	2	155.0–184.9	7	165.0–204.9	21	175.0–224.9	57
175.0–199.9	15	185.0–214.9	31	205.0–244.9	85	225.0–274.9	147
200.0–224.9	42	215.0–244.9	69	245.0–284.9	126	275.0–324.9	124
225.0–249.9	69	245.0–274.9	99	285.0–324.9	97	325.0–374.9	60
250.0–274.9	78	275.0–304.9	75	325.0–364.9	54	375.0–424.9	7
275.0–299.9	67	305.0–334.9	68	365.0–404.9	12	425.0–474.9	1
300.0–324.9	57	335.0–364.9	35	405.0–444.9	1		
325.0–349.9	42	365.0–394.9	10	445.0–484.9	1		
350.0–374.9	18	395.0–424.9	3				
375.0–399.9	5	425.0–454.9	1				
400.0–424.9	2						
425.0–449.9	1						
MEAN	275.4375	276.275		276.400		275.500	
MEDIAN	272.436	272.576		273.889		272.279	
EMPIRICAL MODE	262.500	260.000		265.000		250.000	
APPROXIMATE MODE (CALCULATED BY FIRST METHOD)	266.433	265.178		268.867		265.837	
APPROXIMATE MODE (CALCULATED BY SECOND METHOD)	262.325	260.630		266.320		259.250	

the remaining 200 plots it is 277.250. Substituting these values in the equation above we have

$$M=\frac{(200\times 273.625)+(200\times 277.250)}{400}=275.4375$$

This is the same value as obtained when all of the 400 individuals are grouped in one distribution, using the same class interval.

Another illustration may be given. The average height of a certain variety of oats was obtained for each of three years and was found to be 76.110, 70.840, and 75.200 centimeters, respectively. There were 500 plants measured the first year and 400 in each of the other two years. By the above formula

$$M=\frac{(500\times 76.110)+(400\times 70.840)+(400\times 75.200)}{1300}=74.208$$

This formula indicates that it is possible to combine the means from different frequency distributions of similar material, but it is well to point out that means should not be averaged unless properly weighted, that is weighting by the number of individuals studied. For example, in the case of the three means cited in the last illustration we have from the formula a general mean of 74.208, while if we averaged the three means without weighting, the average would be 74.050.

Suppose we take another example which illustrates a typical case where mistakes are made by obtaining the mean incorrectly. The illustration assumes the yields of rice obtained from farms of different sizes.

NUMBER OF MOW	TOTAL YIELD, CATTIES	AVERAGE YIELD, CATTIES
10	1300	130
50	7500	150
80	16000	200

We find the average yields for the three different areas to be 130, 150, and 200 catties. Now if these averages are summed and divided by 3 we have an average yield of 160 catties per mow.

This is the *wrong* way to treat these data. The *correct* way is to weight in accordance with the above formula, and we have

$$M=\frac{(10\times 130)+(50\times 150)+(80\times 200)}{10+50+80}=177.14$$

From this it is clear that the correct average is 177.14 catties per mow. It should be kept in mind that for correct results averages should never be combined without proper weighting. This is true whether the individual averages are expressed in units of measurement or on a percentage basis.

Median. The next constant of position to be considered is the median. The median may be considered in this light. Suppose that a number of men were arranged according to height, beginning with the shortest and arranging in order to the tallest; then, if there were an odd number the middle one could be measured and this height would represent the median value. If there were an even number the median would be taken as the average of the two individuals at the middle. It is evident that the median may be determined without knowing the exact measurements of the individual items. With the men grouped as above we do not need to know the individual height of any man except the middle one, if the number is odd, or of the two men nearest the middle if the number is even. There would be no change in the median even if a dwarf occurred in the series, or if we should find a giant as the tallest individual in the group. It must also be clear, then, that the median is not necessarily a value that is found in the series. If we have an odd number then the median does represent a measurement found in our series. In the example cited it is the height of the middle man. However, if we have an even number the median value does not correspond with any individual in the whole series.

Considered from the standpoint of a frequency distribution and a plotted curve, the median is the point on the x axis above and below which 50 per cent of the individuals lie. From the foregoing it is apparent that the calculation of the median ought not to be a very difficult process. An example will show how readily it is determined, using the distribution given in Table 12 repeated here as Table 18.

TABLE 18

METHOD OF CALCULATING THE MEDIAN

CLASS	*f*	
125.0–149.9	2	
150.0–174.9	2	
175.0–199.9	15	
200.0–224.9	42	
225.0–249.9	69	130
250.0–274.9	78	
275.0–299.9	67	
300.0–324.9	57	
325.0–349.9	42	
350.0–374.9	18	
375.0–399.9	5	
400.0–424.9	2	
425.0–449.9	1	
	$N = 400$	

$$Mi = 250.0 + \frac{25\,(400/2 - 130)}{78} = 272.436$$

The point sought is the one above and below which one-half (200) of the individuals lie. The first step is to add the frequencies, beginning with the lowest class, until 200 are obtained. Adding all those up to and including class 225.0–249.9 gives 130 individuals. In the next class there are 78 individuals, and there are needed 200–130, or 70, individuals to complete the number desired (200). Now the 78 individuals of class 250.0–274.9 are treated as if they were uniformly distributed throughout the class. With the usual types of frequency distributions this treatment does not introduce any serious error. The question now arises as to how far beyond the lower limit of this class we must proceed to obtain the 70 individuals necessary to complete one-half of the total population. There are 78 individuals altogether in the class, and as we need 70 to make up the total of 200 we must take 70/78 of the class, so we divide 70 by 78, obtaining .89744 as the result. This means that .89744 of the class must be included to obtain the 70 individuals needed. Since the class interval is 25 grams, we must multiply .89744 by 25, giving 22.436 as the distance above the lower limit of class 250.0–274.9 that the median falls. The median will then be 250.0+22.436, or 272.436 grams. This may be put in a formula as follows:

Let Mi = median
L = lower limit of class in which median lies
a = the number of individuals up to the class in which the median lies
b = the number of individuals in the class in which the median lies
i = class interval
N = total number of individuals in the population

Then

$$Mi = L + \frac{i\,(N/2 - a)}{b}$$

Substituting in this formula the values from the example just given, we have

$$Mi = 250.0 + \frac{25\,(400/2 - 130)}{78} = 250.0 + 22.436 = 272.436$$

The mean for this distribution is 275.4375 grams.

Another illustration to show the determination of the median, and the comparison between the median and the mean, is given in Table 19.

TABLE 19

METHOD OF CALCULATING THE MEDIAN. DATA ARE HEIGHTS IN CENTIMETERS OF 400 OAT PLANTS

CLASS	*f*	
45.0–49.9	2	
50.0–54.9	9	
55.0–59.9	20	
60.0–64.9	35	
65.0–69.9	91	157
70.0–74.9	125	
75.0–79.9	91	
80.0–84.9	26	
85.0–89.9	0	
90.0–94.9	1	
	N = 400	

$$Mi = 70.0 + \frac{5\,(400/2 - 157)}{125} = 71.72$$

Mean = 71.00 centimeters

These data are the heights in centimeters of 400 oat plants. As we sum the frequencies up to and including class 65.0–69.9 we have a total of 157. Now the number needed to make exactly 200, one-half of the total population, is 43. This means we must include enough of class 70.0–74.9 to obtain the 43 individuals. As there are 125 individuals in this class we obtain 43/125 × 5 (the class interval). The result is 1.72, which added to the lower limit of the class (70.0) gives a median value of 71.72 centimeters. The mean for this distribution is 71.00 centimeters.

Mode. As stated earlier, there are two types of mode that may be considered. An observation of the different frequency distributions that have been given in this and previous chapters indicates that there is usually one class that has more individuals than any other. This is true of most frequency distributions, especially if the population is large enough to give a fairly satisfactory distribution. The class that has the largest number of individuals is called the modal class, and the mode is designated by either the class limits, or, better, by the mid-point of this class. This is called the empirical mode. The other type of mode is known as the theoretical mode, which is the maximum ordinate of the theoretical curve which most nearly fits the observations. The empirical mode gives a general idea of type without any calculation, since, when the frequency distribution is completed, the empirical mode is determined by the class having the greatest frequency. This empirical mode has its limitations, since it is readily affected by methods of grouping. It is, however, a convenient constant denoting position and may be compared with other modes of similar material obtained from other distributions where the class limits have been the same. The theoretical mode, since it is the maximum ordinate of a theoretical curve, requires special calculation and for this reason is not so easily determined by one unfamiliar with the steps of curve fitting.

That the mode is affected by methods of grouping is shown by referring to the frequency distributions in Table 17, where the same individuals as in Table 12 have been grouped into classes with class intervals of 30, 40, and 50 grams. It will be noted that with the different classifications the empirical modes are as follows:

CLASS INTERVAL 25 GRAMS	CLASS INTERVAL 30 GRAMS	CLASS INTERVAL 40 GRAMS	CLASS INTERVAL 50 GRAMS
Mode = 262.500	260.000	265.000	250.000

It is seen that the empirical mode varies from 250.000 grams to 265.000 grams according to the different methods of grouping. It has been shown that the mean is very little affected by the different classifications.

As already stated, the exact, or theoretical mode, is difficult to calculate as it is the maximum ordinate of the theoretical curve that best fits the distribution. The following relationship gives an approximate value for the mode which for most distributions agrees fairly closely with the theoretical mode.

$$\text{Mode} = \text{Mean} - 3\,(\text{Mean} - \text{Median})$$

or

$$\text{Mode} - \text{Median} = 2\,(\text{Median} - \text{Mean})$$

That is, there is a general relationship between these three constants. The median is about one-third of the distance from the mean toward the mode, or the median lies between the mean and the mode but is nearer the mean than the mode. Using this formula and calculating the modes for the four distributions in Table 20 we have

CLASS INTERVAL 25 GRAMS	CLASS INTERVAL 30 GRAMS	CLASS INTERVAL 40 GRAMS	CLASS INTERVAL 50 GRAMS
Mode = 266.433	265.178	268.867	265.837

While these approximate modes are affected by grouping, there is not the fluctuation between the different approximate modes as was found with the empirical modes. It is evident from the formula that the mode thus calculated includes all of the individuals since both the mean and the median are involved, while the empirical mode is dependent entirely on the class with the greatest frequency. That the mode as calculated by this formula agrees fairly closely with the theoretical mode obtained from the theoretical curve is evident from the following frequency distribution, which gives the total yield in grams of 400 oat plants.

Total Yield of Plant, Grams	f
0.0–0.9	3
1.0–1.9	50
2.0–2.9	106
3.0–3.9	109
4.0–4.9	80
5.0–5.9	42
6.0–6.9	7
7.0–7.9	2
8.0–8.9	1
	400

The empirical mode is 3.500 grams; the theoretical mode 3.1505 grams; and the mode calculated by the above formula 3.2133 grams.

Yule has given several comparisons between the approximate mode, which is the one calculated by the above formula, and the true mode, as follows:

COMPARISON OF THE APPROXIMATE AND TRUE MODES FOR FIVE DISTRIBUTIONS OF PAUPERISM IN THE UNIONS OF ENGLAND AND WALES

Year	Approximate Mode	True Mode
1850	5.767	5.815
1860	4.610	4.657
1870	5.238	5.038
1881	3.217	3.240
1891	3.007	2.987

Another method for obtaining the approximate theoretical mode has been given by Mills. By this method we have

L = lower limit of modal class
f_1 = frequency of class next below modal class in value
f_2 = frequency of class next above modal class in value
i = class interval

$$Mode = L + \frac{f_2}{f_2 + f_1} \times i$$

This may also be changed as follows and used for checking the calculation:

Let L_1 = upper limit of modal class

Then

$$Mode = L_1 - \frac{f_1}{f_2 + f_1} \times i$$

When this method is used it is evident that there is a closer agreement between the empirical mode and the approximate theoretical mode thus calculated. This method does not include the whole distribution as only three classes, the middle class and the class below and above the middle class, are involved. The mode calculated by this method is dependent on the distribution in these three classes and is not at all affected by the remaining classes of the frequency distribution.

The following summary shows clearly the effect of grouping on the mean, median, empirical mode, and the calculated approximate mode for the data appearing in Table 12.

	CLASS INTERVAL 25 GRAMS	CLASS INTERVAL 30 GRAMS	CLASS INTERVAL 40 GRAMS	CLASS INTERVAL 50 GRAMS
MEAN	275.4375	276.275	276.400	275.500
MEDIAN	272.436	272.576	273.889	272.279
EMPIRICAL MODE	262.500	260.000	265.000	250.000
APPROXIMATE MODE (CALCULATED BY FIRST METHOD)	266.433	265.178	268.867	265.837
APPROXIMATE MODE (CALCULATED BY SECOND METHOD)	262.325	260.630	266.320	259.250

In Chapter II the importance of having sufficient numbers for the population was stressed. This is made more evident now by comparing the constants of the distributions in Table 20, where we have used the material appearing in Table 12. In the first distribution we have only 100 of the individuals; in the next two we have 200 in each distribution; and finally we have the distribution for the entire 400 individuals. In the first distribution it is noted that there is no well defined empirical mode, as class 300.0–324.9 has 20 individuals and class 225.0–249.9 has 19 individuals, while the two classes in between have 17 and 16 individuals, respectively. For the second distribution with 200 individuals we see that there is a tendency for more individuals to fall in the class 225.0–249.9, and thus we have the modal class at this point. Taking the next distribution we find that the empirical mode has shifted again to class 250.0–274.9, and for the grouping of the entire 400 individuals we find that the empirical mode remains in class 250.0–274.9. We note also that while the mean does not fluctuate so much for these four distributions yet it does vary from 273.625 to 277.250. The median ranges from 271.094 to 275.000.

TABLE 20

EFFECT OF NUMBER OF INDIVIDUALS ON MEAN, MEDIAN, AND MODE

CLASS	FIRST 100 INDIVIDUALS	FIRST 200 INDIVIDUALS	LAST 200 INDIVIDUALS	TOTAL 400 INDIVIDUALS
125.0–149.9	1	2		2
150.0–174.9	0	0	2	2
175.0–199.9	2	7	8	15
200.0–224.9	11	25	17	42
225.0–249.9	19	39	30	69
250.0–274.9	17	32	46	78
275.0–299.9	16	33	34	67
300.0–324.9	20	29	28	57
325.0–349.9	9	22	20	42
350.0–374.9	2	6	12	18
375.0–399.9	2	3	2	5
400.0–424.9	0	1	1	2
425.0–449.9	1	1		1
MEAN	276.500	273.625	277.250	275.4375
MEDIAN	275.000	271.094	273.370	272.436
EMPIRICAL MODE	312.500	237.500	262.500	262.500

From this illustration it is clear that for the best results in statistical analysis one should have a fairly large population. Usually 100 or 200 individuals are insufficient and it is better to have 400 or more individuals. There are, however, occasions when it is impossible to have a large number in the population and under such conditions one should be careful in drawing his conclusions and use all the safeguards that will be pointed out in later chapters to qualify any deductions he may make.

The three constants of position—arithmetic mean, median, and mode—may now be compared as to their usefulness. In the first place it may be stated that the arithmetic mean is more useful generally than the other two constants. It is more widely understood and in its determination all of the individuals are included and each one affects the value of the arithmetic mean.

The median gives an idea of type and is more quickly determined than the mean. The values of the individuals, except those at the middle or near the middle, do not affect the location of the median. That is, the extreme individuals either below or above the median, no matter what their individual measurements may be, do not affect the location of the median.

The mode is also readily determined, the empirical mode being at once evident from the frequency distribution and the approximate mode being easily calculated. The mode also gives an idea of type.

For a perfect symmetrical distribution these three constants will be located at the same point, but in the asymmetrical distributions so frequently found they differ from each other in value. The question naturally arises as to which one should be used to represent position or type. In answer to this it may be stated that no one of these constants can be recommended to be used at all times in preference to the others. As has already been indicated, the mean is more generally useful, but there are occasions when it may not represent the actual condition as accurately as will the median or mode. For example, if one were studying the incomes of the male members of a certain community, say a Chinese village, one might find the following conditions. It may be possible that in the village there is one rather well-to-do individual while the majority of the men are laborers or farmers, and the incomes might be distributed in accordance with the following frequency distribution.

Income Per Year, Dollars	f
80.0– 99.9	8
100.0– 119.9	13
120.0– 139.9	16
140.0– 159.9	22
160.0– 179.9	17
180.0– 199.9	12
200.0– 219.9	11
6000.0–6019.9	1
	100

From this distribution it is evident that 22 of the 100 individuals have incomes around $150.00, while the one well-to-do individual has an income of $6010.00. Now in this case if we should use the arithmetic mean to represent the incomes we would have an average income of $210.20, but it is clear that the majority of the men receive incomes less than this amount. In this instance the arithmetic mean does not give us the best idea of type. Here the mode will give us a better idea, as the empirical mode is at $150.000 and the approximate mode calculated by the second method referred to above is $150.303. The median, which is $151.818, is also recommended in this case to represent the actual condition.

Thus, for certain kinds of distributions we may find that the mean fails to give as true an idea of the position or type as does one of the other constants. It is necessary that one should become thoroughly familiar with the uses and limitations of these three types of average so that he may know which will give the best expression of the condition in which he is interested or which he wishes to emphasize. Often with social or economic studies the mode or median may be a better constant than the mean.

While discussing averages it may be important to point out the usefulness of the moving average. The moving average may be used to represent the general trend or tendency for results that are recurring over a long period of time, or where it is desired to reduce or smooth the effects of the slight fluctuations that may occur from year to year. An example will make this clear.

TABLE 21

METHOD OF CALCULATING MOVING AVERAGE. DATA FROM UNIVERSITY OF ILLINOIS AGRICULTURAL EXPERIMENT STATION, SHOWING RESULT OF SELECTING MAIZE FOR HIGH OIL CONTENT

YEAR	AVERAGE PERCENTAGE OIL IN CROP EACH GENERATION	3-YEAR MOVING AVERAGE	5-YEAR MOVING AVERAGE
1896	4.70		
1897	4.73	4.86	
1898	5.15	5.17	5.27
1899	5.64	5.64	5.55
1900	6.12	5.95	5.88
1901	6.09	6.21	6.15
1902	6.41	6.33	6.42
1903	6.50	6.63	6.65
1904	6.97	6.92	6.91
1905	7.29	7.21	7.11
1906	7.37	7.36	7.25
1907	7.43	7.33	7.27
1908	7.19	7.22	7.35
1909	7.05	7.32	7.38
1910	7.72	7.43	7.43
1911	7.51	7.64	7.63
1912	7.70	7.79	7.87
1913	8.15	8.05	8.02
1914	8.29	8.30	8.22
1915	8.46	8.42	8.39
1916	8.50	8.50	8.63
1917	8.53	8.79	8.78
1918	9.35	8.98	8.94
1919	9.05	9.23	9.23
1920	9.28	9.42	9.50
1921	9.94	9.69	9.64
1922	9.86	9.96	9.50
1923	10.08	9.93	9.99
1924	9.86	10.05	
1925	10.21		

In Table 21 we have the results by individual years of selecting corn for high oil content. These may be combined in a three-year or a five-year moving average. For the three-year moving average one may center the first result on the year 1897, and thus he would average the result of the previous year and the result of the following year with the result for 1897, and obtain a moving average of 4.86. For the year 1898 he would use the results for 1897, 1898, and 1899.

For the five-year moving average he would proceed as follows. He might center his first result on the year 1898, and for this would average the results of the two previous years and the two succeeding years with the result for 1898. He would continue through the series in this manner.

In certain instances it may be found desirable to follow some system of weighting when obtaining the moving average. That is, there may be certain reasons why the results of one year or of a particular condition may be more important than others, and in such cases one may weight the results by making one result say twice as valuable as another. This is done only in special cases and is not a necessary part of the calculation of moving averages.

Geometric Mean. An illustration of the calculation of the geometric mean is given in Table 22, using the data appearing in Table 12.

TABLE 22

METHOD OF CALCULATING THE GEOMETRIC MEAN

CENTER	LOGS CENTER	POWERS	PRODUCTS OF LOGS AND POWERS
137.5	2.1383	2	4.2766
162.5	2.2109	2	4.4218
187.5	2.2730	15	34.0950
212.5	2.3274	42	97.7508
237.5	2.3757	69	163.9233
262.5	2.4191	78	188.6898
287.5	2.4586	67	164.7262
312.5	2.4948	57	142.2036
337.5	2.5283	42	106.1886
362.5	2.5593	18	46.0674
387.5	2.5883	5	12.9415
412.5	2.6154	2	5.2308
437.5	2.6410	1	2.6410
		400	973.1564

Log 973.1564/400 = 2.4329
Natural number of log 2.4329 = 270.96
Geometric mean = 270.96

This constant is not used very often, and especially is this true in connection with frequency curves and general statistical analysis. The geometric mean is the antilogarithm of the mean of the logarithms of the numbers. Secrist gives the following formula for determining the geometric mean:

$$\text{Geometric mean} = \sqrt[n]{p_1 \times p_2 \times p_3 \times \ldots p_n}$$

(p_1, p_2, etc., referring to the values of the different items, and n to the number of items)

The steps in the calculation of the geometric mean are as follows. First the logarithms of the mid-points (center) of the classes are obtained and these are multiplied by the frequencies (powers), giving the products in the last column. These products are then summed and divided by the total number (400), giving the logarithm (2.4329) for the geometric mean. The antilogarithm (270.96) of this value is the geometric mean.

The geometric mean will always be less than the arithmetic mean, and it has not come into general use partly on account of the difficulty in calculation but more because of the fact that it "does not possess any simple and obvious properties which render its general nature readily comprehensible." Yule states that the use of the geometric mean finds its simplest application in estimating the numbers of a population midway between two epochs (say two census years) at which the population is known.

A brief comparison may now be made between the more common constants of position

ARITHMETIC MEAN	MEDIAN	MODE
Easy to calculate when all the measurements of the individuals can be summed and divided by their number.	Easy to determine when all the individuals may be arranged in order.	Empirical mode easy to determine but readily affected by methods of grouping. Theoretical mode cannot be determined exactly without fitting a theoretical curve to the distribution. May be determined approximately.

ARITHMETIC MEAN	MEDIAN	MODE
Is affected by the measurements of all the individuals. Since this is so it is sometimes too greatly affected by abnormal individuals or individuals removed at some distance from the central tendency of the distribution.	Is determined merely by its position in the distribution and is not affected by abnormal individuals.	Is not so affected by abnormal individuals and is an important constant of position because it is located where the frequency is greatest.
May be treated algebraically, which is one of its important characteristics.	Cannot be treated algebraically, that is the median of two or three distributions cannot be determined from the median values of the individual distributions.	Is unsuited for algebraical work for, like the median, the mode of several distributions cannot be determined by noting the modes of the individual distributions.

CHAPTER V

CONSTANTS OF DEVIATION OR DISPERSION

When the type of a frequency distribution, as defined by the constants of position, has been determined, one is interested in knowing of the spread, or deviation from this type, that the material exhibits. On first consideration it might seem that if one had the type definitely denoted by the mean, then the range of the distribution might tell sufficiently accurately what it is needed to know regarding the spread or variability of the material. However, it will soon be evident that such is not the case. For example, in Table 12 (Chapter IV), giving the yields of soy beans, the range as now given is 325 grams, but if the one individual in the last class were eliminated the range would be only 300 grams. Thus by eliminating one individual out of a population of 400 the range would be decreased by 25 grams. The mean would be only slightly affected, as after eliminating the individual in the last class it would be 275.0313, while the mean for the 400 individuals is 275.4375 grams. Suppose again that an individual had been found that measured only 100 grams, the range would have been increased by 25 grams by the addition of only one individual. When considering the range alone, or the range and the mean together, one does not obtain any definite idea regarding the spread of the material, for these alone do not give an accurate idea of the form of the distribution. Two distributions may be obtained with the same range, yet in one the majority of the population may be grouped closely about the mean while in the other the grouping may spread out considerably from the mean. It is clear, then, that the range is not a good measure of the variability of a group of individuals.

What is needed is a constant for measuring dispersion that satisfies the conditions that were laid down for the constants of position. The best constant of dispersion is one that considers the values of all of the individuals, one that may be treated algebraically, and one that may be useful in many calculations that are based directly or indirectly on the deviation. The constants commonly used are the *average deviation*, the *variance* or its square root the *standard deviation*, and the *quartile deviation* or semi-interquartile range. Those most generally useful are the variance and its square root, the standard deviation. In general, for the analysis of frequency distributions the standard deviation is most commonly used. The variance, however, is especially valuable, as will be seen in later discussions. The constants of position will be considered in the order named.

Average Deviation. The average deviation, or mean deviation, is the average of all deviations from the mean of the series. It is expressed in the same unit of measurement that has been used in recording the items.

When the number of individuals is small the average deviation may be obtained without grouping, as illustrated in Table 23. The formula for the average deviation ($A.D.$) in such cases is

$$A.D. = \frac{\Sigma + D}{N}$$

The plus sign, when used in this and other formulas, means that the operation is completed without regard to sign. With the data in Table 23 the deviation of each individual from the mean (297.35) is determined, and these deviations are summed without regard to sign. The total summation (677.10) is divided by the number of individuals (20), giving the average deviation (33.855). Substituting these values in the formula

$$A.D. = \frac{677.10}{20} = 33.855$$

TABLE 23

METHOD OF CALCULATING THE AVERAGE DEVIATION FROM UNGROUPED MATERIAL. DATA ARE YIELDS IN GRAMS OF SOY BEAN ROWS

YIELD IN GRAMS	D
230	−67.35
291	− 6.35
290	− 7.35
273	−24.35
292	− 5.35
297	− .35
256	−41.35
246	−51.35
258	−39.35
312	14.65
448	150.65
328	30.65
274	−23.35
309	11.65
295	− 2.35
375	77.65
336	38.65
312	14.65
279	−18.35
246	−51.35
5947	$\Sigma + D = 677.10$

$N = 20$

$M = 297.35$

$\Sigma + D$ (without regard to sign) $= 677.10$

$$A.D. = \frac{677.10}{20} = 33.855 \text{ grams}$$

Usually one is dealing with a large number of individuals and, as was found with the constants of position, it is more convenient to obtain the average deviation from a grouped distribution. The method of obtaining the average deviation from a grouped distribution is illustrated in Table 24.

Table 24

Method of Calculating the Average Deviation from the Mean

Class	V	f	$(V-M)$ D	fD
125.0–149.9	137.5	2	–137.9375	–275.8750
150.0–174.9	162.5	2	–112.9375	–225.8750
175.0–199.9	187.5	15	– 87.9375	–1319.0625
200.0–224.9	212.5	42	– 62.9375	–2643.3750
225.0–249.9	237.5	69	– 37.9375	–2617.6875
250.0–274.9	262.5	78	– 12.9375	–1009.1250
275.0–299.9	287.5	67	12.0625	808.1875
300.0–324.9	312.5	57	37.0625	2112.5625
325.0–349.9	337.5	42	62.0625	2606.6250
350.0–374.9	362.5	18	87.0625	1567.1250
375.0–399.9	387.5	5	112.0625	560.3125
400.0–424.9	412.5	2	137.0625	274.1250
425.0–449.9	437.5	1	162.0625	162.0625
		$N=400$		$\Sigma+fD=16182.0000$

$M=275.4375$ $\quad A.D.=\frac{16182.0000}{400}=40.455$

The process consists of taking the deviation of each class from the mean (275.4375 grams in this case), multiplying each deviation by its respective frequency, and summing without regard to sign. This sum is then divided by the total number, N, and the result is the average deviation. The formula is

$$A.D.=\frac{\Sigma+fD}{N}$$

Substituting the values from Table 24 in the above formula

$$A.D.=\frac{16182.0000}{400}=40.455 \text{ grams}$$

It is to be noted that the average deviation is given in grams, or the unit used in obtaining the data.

Sometimes it may be convenient to compute the average deviation from an assumed mean, but care must be taken when so doing. The method of calculating the average deviation from an assumed mean is illustrated in Table 25, with the same data as were used in Table 24.

TABLE 25

METHOD OF CALCULATING THE AVERAGE DEVIATION FROM AN ASSUMED MEAN

Assumed Mean $(G) = 287.5$

CLASS	V	f	D	fD
125.0–149.9	137.5	2	−150	− 300
150.0–174.9	162.5	2	−125	− 250
175.0–199.9	187.5	15	−100	−1500
200.0–224.9	212.5	42	− 75	−3150
225.0–249.9	237.5	69	− 50	−3450
250.0–274.9	262.5	78 (208)	− 25	−1950 −10600
275.0–299.9	287.5	67	0	
300.0–324.9	312.5	57	25	1425
325.0–349.9	337.5	42	50	2100
350.0–374.9	362.5	18	75	1350
375.0–399.9	387.5	5	100	500
400.0–424.9	412.5	2	125	250
425.0–449.9	437.5	1 (192)	150	150 5775
		$N = 400$		$\Sigma = 16375$

$$c = \frac{4825}{400} = 12.0625$$

To obtain *A.D.* we have

$$16375 - (208 \times 12.0625) + (192 \times 12.0625) =$$

$$16375 - 2509.000 + 2316.000 = 16182$$

$$A.D. = \frac{16182}{400} = 40.455$$

When using this method for calculating the average deviation from an assumed mean, the several deviations from the assumed mean are multiplied by their respective frequencies. There will result a series of negative values and a series of positive values. The difference between these two values will be used in making the correction for the average deviation, and will be referred to as the total difference or the total error. This total difference, divided by the number in the population, will give the average error. If the assumed mean should be the same as the true mean, then it is evident that there will be as large a minus value as a plus value, and the algebraic sum will be 0. In actual practice this very seldom happens.

With the data in Table 25 the average error is used in making the proper correction in the following way. The assumed mean is 287.5 and the deviations (D) of the mid-points of the several classes from this assumed mean are obtained and multiplied by their respective frequencies, giving the values in column fD. These values are summed without regard to sign and the total is 16375. The difference between the plus and minus values, 10600 and 5775, is 4825, the amount of the total error resulting from working from an assumed mean. This amount, 4825, divided by the number in the population, 400, equals 12.0625, the average error. When correcting, the sign of this average error need not be considered.

Now the deviations of the frequencies of the lower part of the distribution, that is those below the class of the assumed mean, which total 208, are too large since they were calculated from 287.5 rather than from 275.4375, the true mean. That is, since the assumed mean was higher than the true mean the frequencies of these individuals in the lower part of the distribution are too large, and similarly, those in the upper part of the distribution, 192, are too small. It is necessary to make a correction for the total, 16375, on account of this fact. Multiplying the average error, 12.0625, by the frequencies 208 and 192 gives the correction for the total deviations, and the total of the values in column fD must be corrected by this amount. We have

$$16375-(208\times 12.0625)+(192\times 12.0625)=16182$$

This value, 16182, divided by 400, gives 40.455, the same average deviation as obtained by the longer method.

This process may be shortened somewhat. Since the frequencies of 208 individuals are too large and the frequencies of 192 individuals are too small, we may correct by taking the difference between these two frequencies, keeping in mind which group is too large and which is too small. The difference between 208 and 192 is 16, and since the larger number, 208, is from the group in which the frequencies are too large it means the final value must be corrected by subtracting 16 times the average error, 12.0625. The product is 193.000, and subtracting this value from 16375 gives the corrected value which, when divided by the number of individuals, 400, gives the average deviation, 40.455. This is the same value as obtained before. In making the correction in this way it is always necessary to keep in mind which group of frequencies is too high and which is too low, so that when taking the difference between the frequencies one will know whether to subtract or add the correction.

If the assumed mean falls outside of the group in which the true mean lies, then the above method needs modification. It is usually more convenient and avoids possibility of error if the deviations are obtained from the true mean and the average deviation determined as in Table 24. It is evident that when the deviations are obtained from the true mean the sum of the negative values is equal to the sum of the positive values. Since this is true it is possible to obtain the average deviation by summing either the negative or positive values, using the formula

$$A.D.=2/N\ (\Sigma fD_x)$$

where ΣfD_x = summation of either negative or positive values. Summing the positive values in Table 24 we have 8091 and substituting this value in the formula

$$A.D.=2/400\times 8091=40.455$$

The same result is obtained by summing the negative values, disregarding the sign.

From this relation the average deviation of a frequency distribution may be obtained from an assumed mean.

Let N = number of individuals in a population
n_a = number of individuals above the true mean
n_b = number of individuals below the true mean
ΣfD_a = sum of the product values of the n_a variates from the assumed mean
ΣfD_b = sum of the product values of the n_b variates from the assumed mean
c = correction or the difference between the true mean and the assumed mean

When the *true mean* is *larger* than the assumed mean the formula is

$$A.D=2/N\ (\Sigma fD_a-cn_a)$$

When the *true mean* is *smaller* than the assumed mean the formula is

$$A.D.=2/N\ (\Sigma fD_b-cn_b)$$

With the data in Table 25 the true mean is smaller than the assumed mean. The sum of the frequencies below the true mean is 208 and the sum of their product values is 10600. The correction, or the difference between the true mean and the assumed mean, is 12.0625. Substituting these values in the formula to be used when the true mean is lower than the assumed mean

$$A.D.=2/400\ [10600-(208\times 12.0625)]=8091/200=40.455$$

With the data in Table 26 the true mean is higher than the assumed mean.

TABLE 26

METHOD OF CALCULATING THE AVERAGE DEVIATION FROM AN ASSUMED MEAN OUTSIDE THE CLASS IN WHICH THE TRUE MEAN IS FOUND

Assumed Mean $(G)=262.5$

CLASS	V	f	D	fD
125.0–149.9	137.5	2	−125	−250
150.0–174.9	162.5	2	−100	−200
175.0–199.9	187.5	15	−75	−1125
200.0–224.9	212.5	42	−50	−2100
225.0–249.9	237.5	69	−25	−1725 −5400
250.0–274.9	262.5	78 (208)	0	
275.0–299.9	287.5	67	25	1675
300.0–324.9	312.5	57	50	2850
325.0–349.9	337.5	42	75	3150
350.0–374.9	362.5	18	100	1800
375.0–399.9	387.5	5	125	625
400.0–424.9	412.5	2	150	300
425.0–449.9	437.5	1 (192)	175	175 10575
		$N=400$		$\Sigma=15975$

$$c=\frac{5175}{400}=12.9375$$

$$A.D.=\frac{2}{400}\,[10575-(192\times 12.9375)]=\frac{8091}{200}=40.455$$

There are 192 individuals above the true mean and the sum of their product values is 10575. Substituting these values in the formula to be used when the true mean is higher than the assumed mean

$$A.D. = 2/400\,[10575 - (192 \times 12.9375)] = 8091/200 = 40.455$$

It may be pointed out that if desired the average deviation may be calculated by the unity-step method, as was illustrated in determining the mean. Often such a procedure will save time in calculation.

In the foregoing illustrations the calculations have been made from the mean. It may be pointed out that the average deviation can also be determined from the median. If the median is used rather than the mean the average deviation will be smaller, as the average deviation has a minimum value when calculated from the median. For example, the average deviation for the distribution in Table 26 is 40.335 when calculated from the median. This is slightly less than 40.455, the average deviation obtained by calculating from the mean.

The average deviation has the advantage of being easy to calculate and may be conveniently used in some cases where a measure of variability is wanted quickly. It does not lend itself to algebraic treatment and is not so useful a constant as the standard deviation.

Standard Deviation. As has been suggested, the variance, which is the arithmetic mean of the squares of all deviations measured from the mean, has a useful place in statistics. The square root of the variance, or standard deviation, is used more commonly as a constant of dispersion in the general study of frequency distributions. Yule defines the standard deviation as the square root of the arithmetic mean of the squares of all deviations, deviations being measured from the arithmetic mean of the observations. The standard deviation is designated by *St. Dev.*, or *S. D.*, and very often by the small Greek letter *Sigma* (σ).

While the average deviation is obtained from the first moment, $\Sigma D/N$, the standard deviation is obtained from the second moment, $\Sigma D^2/N$. The formula is

$$S.D. = \sqrt{\Sigma D^2/N}$$

The moments here refer to the summation of all the deviations about the mean divided by the number of individuals. When there are only a few individuals under observation the deviation of each one from the mean may be obtained, then squared, and the total number summed. This summation divided by the number of individuals gives the variance, and the square root of this variance is the standard deviation. This method of calculating the standard deviation from ungrouped material is shown in Table 27.

TABLE 27

METHOD OF CALCULATING THE STANDARD DEVIATION FROM UNGROUPED MATERIAL. DATA ARE YIELDS IN GRAMS OF SOY BEAN ROWS

YIELD IN GRAMS	D	D^2
230	−67.35	4536.0225
291	− 6.35	40.3225
290	− 7.35	54.0225
273	−24.35	592.9225
292	− 5.35	28.6225
297	− .35	.1225
256	−41.35	1709.8225
246	−51.35	2636.8225
258	−39.35	1548.4225
312	14.65	214.6225
448	150.65	22695.4225
328	30.65	939.4225
274	−23.35	545.2225
309	11.65	135.7225
295	− 2.35	5.5225
375	77.65	6029.5225
336	38.65	1493.8225
312	14.65	214.6225
279	−18.35	336.7225
246	−51.35	2636.8225
5947		$\Sigma D^2 = 46394.5500$

$N = 20 \qquad M = 297.35$

$\Sigma D^2 = 46394.5500$

$$S.D. = \sqrt{\frac{46394.5500}{20}}$$

$$= \sqrt{2319.727500}$$

$$= 48.164$$

Usually one is concerned with a large number of individuals, and in such cases it is better to make a frequency distribution and work from the grouped material. This method of calculating the standard deviation from a frequency distribution is shown in Table 28, using the same data as in Table 12.

TABLE 28

METHOD OF CALCULATING THE STANDARD DEVIATION FROM THE TRUE MEAN

CLASS	V	f	$V\text{-}M$	$(V\text{-}M)^2$	$f(V\text{-}M)^2$
125.0–149.9	137.5	2	−137.9375	19026.7539	38053.5078
150.0–174.9	162.5	2	−112.9375	12754.8789	25509.7578
175.0–199.9	187.5	15	− 87.9375	7733.0039	115995.0585
200.0–224.9	212.5	42	− 62.9375	3961.1289	166367.4138
225.0–249.9	237.5	69	− 37.9375	1439.2539	99308.5191
250.0–274.9	262.5	78	− 12.9375	167.3789	13055.5542
275.0–299.9	287.5	67	12.0625	145.5039	9748.7613
300.0–324.9	312.5	57	37.0625	1373.6289	78296.8473
325.0–349.9	337.5	42	62.0625	3851.7539	161773.6638
350.0–374.9	362.5	18	87.0625	7579.8789	136437.8202
375.0–399.9	387.5	5	112.0625	12558.0039	62790.0195
400.0–424.9	412.5	2	137.0625	18786.1289	37572.2578
425.0–449.9	437.5	1	162.0625	26264.2539	26264.2539
		$N=400$		$\Sigma f\,(V\text{-}M)^2=$	971173.4350

$M=275.4375$ $\qquad S.D.=\sqrt{\dfrac{971173.4350}{400}}=\sqrt{2427.933587}=49.274$

The process consists of obtaining the deviations of each class from the mean, or $V\text{-}M$. These deviations are then squared, giving $(V\text{-}M)^2$, and multiplied by their respective frequencies. The values thus obtained are summed, giving $\Sigma f(V\text{-}M)^2$. This summation is divided by the number of individuals and the square root of the result is taken. The formula is

$$S.D=\sqrt{\Sigma f\,(V\text{-}M)^2/N}$$

Substituting the values obtained in Table 28

$$S.D.=\sqrt{971173.4350/400}=\sqrt{2427.933587}=49.274$$

The value of the standard deviation, 49.274, is expressed in grams, the same unit of measurement that was used in obtaining the weight of seed of each plot.

This method of obtaining the standard deviation is somewhat laborious, since one often has to deal with large numbers. A shorter method is illustrated in Table 29, using the same data and following the plan used to determine the mean by working from an assumed mean.

TABLE 29

METHOD OF CALCULATING THE STANDARD DEVIATION BY WORKING FROM AN ASSUMED MEAN

Assumed Mean $(G) = 262.5$

CLASS	V	f	D	fD	fD^2
125.0–149.9	137.5	2	–125	– 250	31250
150.0–174.9	162.5	2	–100	– 200	20000
175.0–199.9	187.5	15	– 75	–1125	84375
200.0–224.9	212.5	42	– 50	–2100	105000
225.0–249.9	237.5	69	– 25	–1725	43125
250.0–274.9	262.5	78	0		
275.0–299.9	287.5	67	25	1675	41875
300.0–324.9	312.5	57	50	2850	142500
325.0–349.9	337.5	42	75	3150	236250
350.0–374.9	362.5	18	100	1800	180000
375.0–399.9	387.5	5	125	625	78125
400.0–424.9	412.5	2	150	300	45000
425.0–449.9	437.5	1	175	175	30625
		$N = 400$		10575	$\Sigma fD^2 = 1038125$
				– 5400	
				$\Sigma fD = 5175$	

$$c = \frac{5175}{400} = 12.9375$$

$$S.D. = \sqrt{\frac{1038125}{400} - (12.9375)^2}$$

$$= \sqrt{2427.933594}$$

$$= 49.274$$

With this method a mean is assumed which may be at any point either within or without the distribution. It makes no difference what value is selected so long as the proper correction is made. Usually it is better to select a value near the middle of the distribution since this will involve smaller values for the deviations than if an extreme value is chosen, and the multiplications will be made more easily. On the other hand, if the mid-point of the first class is chosen one does not have to deal with minus signs, and in some instances and for certain kinds of calculations this may be helpful.

After the mean is assumed, the deviations of the several classes from this mean are set down in column *D*. These deviations are multiplied by their respective frequencies, giving column *fD*, and these values are summed, having regard to the signs. The algebraic sum is obtained and this is divided by *N*, giving *c*, the correction for the mean, or the difference between the true and the assumed mean. As stated in Chapter IV, this value is added to the assumed mean to obtain the true mean. The sign must be observed in obtaining the mean, so that where a negative value is obtained it is really subtracted. This point must be watched carefully.

The next step is to determine fD^2. Since $fD \times D = fD^2$ it is not necessary to write down the column of D^2 values but simply multiply the values in column *D* by their corresponding values in column *fD*. The resulting products are summed, giving ΣfD^2, and divided by *N*, or $\Sigma fD^2/N$. This summation will always be too high when working from an assumed mean, and must be corrected by the value of the correction, *c*. Since, in obtaining the fD^2 values one is working with the second powers, or squares, it is necessary to square *c* before subtracting from $\Sigma fD^2/\mathrm{N}$. The formula for this method of calculating the standard deviation is

$$S.D. = \sqrt{\Sigma fD^2/N - c^2}$$

Substituting the values obtained in Table 29

$$S.D. = \sqrt{1038125/400 - (12.9375)^2} = \sqrt{2427.933594} = 49.274$$

It was stated above that the value of $\Sigma fD^2/N$ will always be higher when working from an assumed mean rather than from the true mean. Perhaps an algebraic proof will make this clear.

Let M = true mean
G = assumed mean
c = correction for the mean
$M = G + c$
V = class value
$V\text{-}M$ = an amount designated by a
$V\text{-}G$ = an amount designated by d

Then

$$V\text{-}M = V-(G+c), \text{ since } M=(G+c), = V\text{-}G\text{-}c$$

By substitution

$$a = d - c$$

Transpose

$$a + c = d$$

Square

$$a^2 + 2ac + c^2 = d^2$$

Multiply by the frequency, f, sum and divide by N

$$\Sigma fa^2/N + \Sigma f\,2ac/N + \Sigma fc^2/N = \Sigma fd^2/N$$

Now $\Sigma f\,2ac/N = 0$, since the sum of the products of the deviations around the mean is 0. In other words, when the deviations are measured from the mean, the sum of the deviations will be 0, since there are as many negative as there are positive values. This can be proved by summing the negative and positive values of the last column of Table 24. The values thus summed are -8091.0000 and +8091.0000.

Therefore

$$\Sigma fa^2/N + \Sigma fc^2/N = \Sigma fd^2/N$$

By transposing, and since

$$\Sigma f = N$$

Then

$$\Sigma fa^2/N = \Sigma fd^2/N - c^2$$

or, the sum of the squares measured from the mean equals the sum of the squares measured from an assumed mean minus c^2, or the square of the correction.

Now, since

$$S.D. = \sqrt{\Sigma fa^2/N}$$

Therefore

$$S.D. = \sqrt{\Sigma fd^2/N - c^2}$$

Thus the deviations measured from an assumed mean are always higher than when measured from the true mean, and c^2 must always be subtracted from $\Sigma fd^2/N$, or as designated above $\Sigma fD^2/N-c^2$.

This fact may also be demonstrated geometrically, as illustrated by Figure 11.

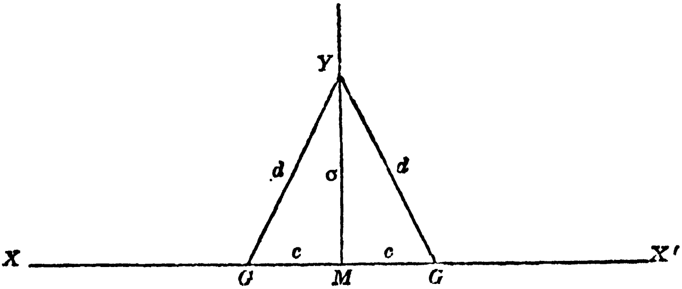


FIG. 11. Illustrating the fact that deviations measured from the mean, M, are less than when measured from any assumed point, G.

Suppose on the horizontal line XX' a perpendicular is erected at M, the mean of a series. Now let a distance MY be marked off equal to the standard deviation, σ. Then let G be any value assumed as the mean of the series and the line YG, or d, be drawn. The difference between M and G is c, the correction for the mean. Now since the line σ is drawn through the mean it is shorter than any other line drawn from XX' to point Y. This may be proved since σ, d, and c are three sides of a right-angled triangle and

$$\sigma^2+c^2=d^2$$

or

$$\sigma^2=d^2-c^2$$

and

$$\sigma=\sqrt{d^2-c^2}$$

The standard deviation is the root-mean-square deviation from the mean and d is the root-mean-square deviation from the assumed origin, G. Thus it follows, as indicated, that the root-mean-square is smallest when calculated from the mean and that when calculated from any other origin the sum of the squares is too high and a correction must be made, depending on the value of c^2.

The calculation of the standard deviation may be shortened still more by the unity-step method, as in the case with the mean. This is done by assuming that the class interval is one unit, as was done in the calculation of the mean by the unity-step method, rather than the actual value. The columns are arranged and the work carried out as indicated in Table 30, using the same distribution as before. The formula for this method is

$$S.D. = \sqrt{\Sigma fD^2/N - c^2} \times ci$$

TABLE 30

METHOD OF CALCULATING THE STANDARD DEVIATION BY THE UNITY-STEP METHOD

Assumed Mean $(G) = 262.5$

CLASS	V	f	D	fD	fD^2
125.0–149.9	137.5	2	–5	–10	50
150.0–174.9	162.5	2	–4	– 8	32
175.0–199.9	187.5	15	–3	–45	135
200.0–224.9	212.5	42	–2	–84	168
225.0–249.9	237.5	69	–1	–69	69
250.0–274.9	262.5	78	0		
275.0–299.9	287.5	67	1	67	67
300.0–324.9	312.5	57	2	114	228
325.0–349.9	337.5	42	3	126	378
350.0–374.9	362.5	18	4	72	288
375.0–399.9	387.5	5	5	25	125
400.0–424.9	412.5	2	6	12	72
425.0–449.9	437.5	1	7	7	49
		$N = 400$		423 –216 $\Sigma fD = 207$	$\Sigma fD^2 = 1661$

$$c = \frac{207}{400} = .5175$$

$$S.D. = \sqrt{\frac{1661}{400} - (.5175)^2} \times 25$$

$$= \sqrt{3.884694} \times 25$$

$$= 49.274$$

The calculations are made on the assumption that the classes differ by unity. After the square root has been obtained it is multiplied by the class interval, if the class interval is any other value than unity, and the result is the true or corrected standard deviation. Substituting the values from Table 30 in the formula

$$S.D. = \sqrt{1661/400 - (.5175)^2} \times 25$$

$$= \sqrt{3.884694} \times 25$$

$$= 49.274$$

The value of the standard deviation, 49.274, is the same as obtained by the longer methods. The unity-step method is much simpler to use, since the numbers are smaller and the calculations may be handled conveniently without the aid of calculating machines.

Another example will illustrate the unity-step method of calculating the standard deviation.

TABLE 31

METHOD OF CALCULATING THE STANDARD DEVIATION BY THE UNITY-STEP METHOD

DATA ARE HEIGHTS IN CENTIMETERS OF 400 OAT PLANTS

Assumed Mean $(G) = 72.5$

CLASS	V	f	D	fD	fD^2
45.0–49.9	47.5	2	–5	–10	50
50.0–54.9	52.5	9	–4	–36	144
55.0–59.9	57.5	20	–3	–60	180
60.0–64.9	62.5	35	–2	–70	140
65.0–69.9	67.5	91	–1	–91	91
70.0–74.9	72.5	125	0		
75.0–79.9	77.5	91	1	91	91
80.0–84.9	82.5	26	2	52	104
85.0–89.9	87.5	0	3	0	0
90.0–94.9	92.5	1	4	4	16
		$N = 400$		–267	$\Sigma fD^2 = 816$
				147	
				$\Sigma fD = -120$	

$$c = \frac{-120}{400} = -.300$$

$$S.D. = \sqrt{\frac{816}{400} - (-.300)^2} \times 5$$

$$= \sqrt{1.950000} \times 5$$

$$= 6.982$$

The various steps are no doubt clear. Substituting in the formula the values obtained in Table 31

$$S.D. = \sqrt{816/400 - (-.300)^2} \times 5$$

$$= \sqrt{1.950000} \times 5$$

$$= 6.982$$

It may be well to emphasize again the importance of observing the signs in the column *fD* and also the calculations from the sum of this column to insure the proper correction for the mean. In practical work it is necessary to retain only sufficient decimals for accuracy. It should be pointed out, however, that when the class interval is large a slight error may result by reading to only three decimals and later obtaining the true mean or standard deviation by multiplying by a large class interval, such as 50 or 100, and sometimes by a smaller amount. Therefore to be correct to three decimals the calculations should be carried far enough so that the constant will be correct to the number of places to be kept in the record.

Before leaving the standard deviation it may be well to point out two other useful formulas that may be used to obtain the standard deviation. They are especially useful when the population is not too large and one may be working from ungrouped data. These methods are illustrated in Table 32.

TABLE 32

ILLUSTRATING SPECIAL METHODS OF DETERMINING STANDARD DEVIATION FROM UNGROUPED DATA

x	x^2
50	2500
42	1764
54	2916
56	3136
47	2209
48	2304
53	2809
42	1764
45	2025
50	2500
51	2601
538	$\Sigma x^2 = 26528$

$N = 11$

First Method:

$$M = \frac{538}{11} = 48.909$$

$$S.D. = \sqrt{\frac{26528}{11} - (48.909)^2}$$
$$= \sqrt{19.546083}$$
$$= \quad 4.421$$

Second Method:

$$S.D. = \sqrt{\frac{11\ (26528) - (538)^2}{(11)^2}}$$
$$= \sqrt{\frac{291808 - 289444}{121}}$$
$$= \sqrt{\frac{2364}{121}}$$
$$= \quad 4.420$$

The first method is to work directly from the values of the observed items, x. These are summed and the mean is obtained. Instead of obtaining the differences between the several items and the mean, the mean is assumed to be 0 and the x values are squared directly and summed. This sum is then divided by the number of individuals, N, and the correction in this case is the mean, M. The formula is

$$S.D. = \sqrt{\Sigma x^2/N - M^2}$$

With the values in Table 32

$$S.D. = \sqrt{26528/11 - (48.909)^2}$$
$$= \sqrt{19.546083}$$
$$= \quad 4.421$$

The standard deviation for this distribution determined in the usual way by obtaining the deviation of each item from the mean and squaring is 4.420, a value very close to the one obtained by the shorter method.

The formula for the second method is

$$S.D. = \sqrt{\frac{N\ (\Sigma x^2) - (\Sigma x)^2}{N^2}}$$

The several items are squared and summed and the total is multiplied by N. From this result is subtracted the square of the total sum of the x values. The remainder is divided by the square of N and the square root extracted. Substituting the values in Table 32

$$\begin{aligned} S.D. &= \sqrt{\frac{11\ (26528) - (538)^2}{(11)^2}} \\ &= \sqrt{\frac{291808 - 289444}{121}} \\ &= \sqrt{\frac{2364}{121}} \\ &= \quad 4.420 \end{aligned}$$

This method is very useful if the population is not too large and has the advantage that if the mean cannot be obtained exactly there is no error introduced in the squares, as is the case when the deviations are obtained from an approximate mean. For example, in this case the mean is 48.909+and for each deviation that would be obtained there would be a slight error. Since the total, Σx, is used and not $\Sigma x/N$, these slight errors are eliminated and at the same time the calculations are simplified and shortened.

The standard deviation measures the spread or variability of the group for which it has been calculated and is, of course, always expressed in the same unit of measurement that has been used to make the record. One of the general empirical rules is that six times the standard deviation will include approximately 99 per cent of all the observations. This holds more generally for frequency distributions that are symmetrical or only slightly asymmetrical. For example, in Table 28, six times the standard deviation, 49.274, is 295.644, and a range of this amount contains nearly all of the 400 individuals. If one-half this amount is added to or subtracted from the mean we have a range from 127.6155 to 423.2595, and this includes practically all the individuals except the one in class 425.0–449.9.

For distributions that are symmetrical or only slightly asymmetrical there is a relation between the average deviation and standard deviation. The average deviation is approximately four-fifths of the standard deviation. The exact relation is shown by

$$A.D. = .7979\ S.D.$$

For the data shown in Table 28, the average deviation is 40.455 and the standard deviation is 49.274. When the average deviation is divided by the standard deviation the quotient is .8210. From the results in Table 31 the average deviation is 5.497 and the standard deviation is 6.982. Dividing the average deviation by the standard deviation the quotient is .7873. While these values differ from the theoretical value of .7979, it should be borne in mind that the exact relationship as stated above holds only in symmetrical distributions.

It has been stated that six times the standard deviation includes about 99 per cent of all the individuals of a distribution. If the average deviation is to be used for this purpose instead of the standard deviation, it must be multiplied by the factor 7.5.

In discussing the mean it was pointed out that the mean of several series taken together could be obtained from the means of the individual series by applying the formula

$$M=\frac{n_1M_1+n_2M_2+n_3M_3,\text{ etc.}}{N}$$

In a similar manner the standard deviation of several series may be obtained.

If M is the mean of all the series together and M_1 and M_2 the means of the individual series, then

$$M_1-M=d_1$$
$$M_2-M=d_2$$

Now the sum of the deviations of each series about the mean, M, is equal to $n_1\ (\sigma_1^2+d_1^2)$ and $n_2\ (\sigma_2^2+d_2^2)$. Therefore, if σ equals the standard deviation for the whole series, we have

$$N\sigma^2=n_1\ (\sigma_1^2+d_1^2)+n_2\ (\sigma_2^2+d_2^2)$$

and

$$\sigma^2=\frac{n_1\ (\sigma_1^2+d_1^2)+n_2\ (\sigma_2^2+d_2^2)}{N}$$

This may at times prove to be a useful formula.

Coefficient of Variability. As has been stated, the standard deviation is an expression measuring the variability of a group of individuals from their mean and is expressed in units of measurement. For this reason one cannot compare directly the variability as measured by the standard deviation of a group whose measurements have been recorded in grams with another whose measurements have been recorded in inches, and the like. When dealing with distributions of similar material where the method of classification has been the same or nearly the same the standard deviation may be used as an index of variability. For example, if the heights of several groups of men of the same race are to be compared and the classification has been handled similarly, then the standard deviation may be used in comparing the variability of the different groups. However, if the several distributions differ in position and units of measurement, then the standard deviation as an index of variability may be limited in its usefulness.

In order, then, to compare variabilities based on different units, a relative measure of variability has been brought into use. It is the coefficient of variability and is expressed

$$C=\frac{S.D.}{M}\times 100$$

In other words, it is the standard deviation expressed as a percentage of the mean. This constant considers both the variability as expressed by the standard deviation and the position of the distribution as expressed by the mean, and therefore gives a constant expressing relative variability.

By the use of this constant the variability of different groups may be compared directly. For example, with the two lots of data presented in Tables 30 and 31, those in Table 30 deal with the yields per plot of soy beans, measured in grams, while the data in Table 31 deal with the heights of oat plants measured in centimeters. In the first case the standard deviation is 49.274, and in the second case it is 6.982. The first standard deviation is several times larger than the second, but we cannot say from this that the yield of soy beans per plot is several times more variable than the height of oat plants. We must express this variability with relation to the mean in each case. In other words, we use the coefficient of variability, and find that the coefficient of variability for the yields of the soy bean plots is 17.889 while that for the heights of the oat plants is 9.834. While the variability of the yields of the soy bean plots is greater than that of the heights of the oat plants, it cannot be said that it is several times higher, as was indicated by the standard deviation, but only twice as high.

In general practice, therefore, it is always better to obtain the coefficient of variability. While both the standard deviation and the coefficient of variability may be used as an index of variability, it should be remembered that when distributions vary in position and when different units of measurement have been used the coefficient of variability is to be preferred. There are cases having a very low mean and a very great spread of the distribution, resulting in a standard deviation higher than the mean value. In such cases the coefficient of variability may be extremely high and care should be used in interpreting the variability constants.

Quartile Deviation. The median has already been explained as that point on the x axis which divides the population into exactly one-half. The quartiles may now be discussed. We have Q_1, Q_2, and Q_3. The first quartile, Q_1, is the point on the x axis below which one-quarter of the values lie. Three-quarters of the values lie above it. The second quartile Q_2 is the same as the median, and Q_3 is the point below which three-fourths of the values lie, with one-fourth greater.

The quartile deviation is determined by the following relation

$$Q=\frac{Q_3-Q_1}{2}$$

The coefficient that is sometimes used corresponding to the coefficient of variability is to divide the quartile deviation, Q, by the median, or $\frac{Q}{Mi}$.

The values of Q_1 and Q_3 may be determined in a manner similar to that described for the median. The data that served for the calculation of the median, in Table 18, are used here.

Total number of individuals $=400$
One-fourth the number of individuals $=100$
Number of individuals up to class 225.0–249.9 $=61$
Difference between 61 and 100 $=39$
Frequency of class 225.0–249.9 $=69$
Class interval $=25$

$$Q_1=225.0+\left(\frac{39}{69}\times 25\right)=239.130$$

In a similar manner Q_3 may be obtained.

Total number of individuals	=400
Three-fourths the number of individuals	=300
Number of individuals up to class 300.0–324.9	=275
Difference between 275 and 300	= 25
Frequency of class 300.0–324.9	= 57
Class interval	= 25

$$Q_3=300.0+\left(\frac{25}{57}\times25\right)=310.965$$

From the values of Q_1 and Q_3 we have

$$Q=\frac{Q_3-Q_1}{2}=\frac{310.965-239.130}{2}=35.9175$$

There is a definite relation between the standard deviation and the quartile deviation for symmetrical curves, and for curves that are only slightly asymmetrical the relation also holds. This relation is expressed as

$$Q=.6745\times S.D.$$

or the quartile deviation is approximately a little more than two-thirds of the standard deviation. The importance of this relationship will be shown in later discussions. The standard deviation for the distribution used above is 49.274 (Table 28). If the quartile deviation, 35.9175, is divided by the standard deviation the result is .7289, while for strictly symmetrical curves it should be .6745.

The three constants of dispersion — average deviation, standard deviation, and quartile deviation — are constants that measure dispersion in the units of measurement that were used in obtaining the original records. These may be considered as measures of absolute variability. The average deviation is not generally used. It is a constant that may be used to give an approximate expression of variability, and the same is true of the quartile deviation. Of the three constants the quartile deviation is the most readily determined, that is with the least calculation, but except for the symmetrical or nearly symmetrical curves it does not express the variability as well as does the standard deviation. The standard deviation is the most useful of the three constants. The coefficient of variability is used to compare different degrees of dispersion when the material has been measured in various units of measurement.

CHAPTER VI

SIMPLE CORRELATION

In the discussion of the constants of position and of dispersion in the previous chapters we have considered the variation of only one character at a time. That is, by the methods given we can determine certain facts regarding the height of men, for example, or again certain facts regarding the weight of these same men. Often, however, the problem concerns how such characters vary together, or whether there is any relationship between two such characters. There are many problems in biological, social, and economic statistics where a study of relationship or association is of greater importance than merely to determine the variation of any one of the factors alone. Not only may we be interested in learning how two characters may be associated, but sometimes the relationships of more than two characters are important. That is, there may be certain causes operating to produce a certain effect and the question arises as to the definite relation between the factors or causes and the final result, and how this relation may be measured.

One is interested to know whether as one character or factor increases there is a corresponding increase, or a decrease, in a second character. For example, if a certain man is found who is 66 inches tall and weighs 165 pounds, what may be the expected weight of another man who may be 68 inches tall? The one man who is 66 inches tall has an average weight of 2½ (165/66) pounds for each inch in height. From this is it to be expected that the man 68 inches tall would add another 2½ pounds for each additional inch in height and therefore weigh 170 pounds? If this does not follow, then it is interesting to know what is the relation between weight and height. For example, in a population of several hundred men how close is the association between weight and height?

There are numerous other problems in relationship where it is desired to determine the measure of such relation. Such questions as the relation between height and yield of wheat plants; between length of head of wheat plants and the number of kernels produced; between the size of farm and amount of income; between weather factors, as rainfall, sunshine, and the like, and the yields of crops; and many other similar problems, frequently arise. We think of characters that are associated as being correlated, and in order to measure the intensity of the association we have the methods which have been derived for the analysis of correlation. With many factors in whose correlation we may be interested we do not find absolute or perfect correlation, nor do we find complete independence. Different factors will show different degrees of correlation, and often the same factors obtained under different conditions will show different degrees of association. Such relationships may be shown by a diagram, as the scatter diagram in Figure 12 giving the relation between the weight and height of men.

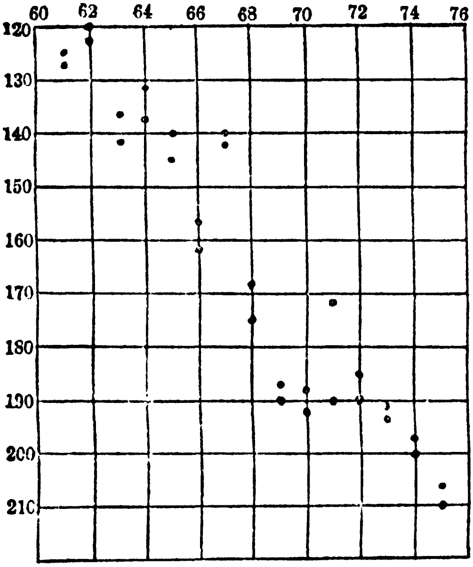


FIG. 12. Scatter diagram showing relation between weight and height of men.

This scatter diagram is made up in the following manner from the data in Table 33, page 92. For example, the second weight in the table is given as 120 pounds and the height of that particular man is given as 62 inches. The dot for this pair of characters

Table 33

Data UsTA in ScatterDiagram Showing Relation Between Weight and Height of Men

Weight in Pounds x	Height in Inches y	Weight in Pounds x	Height in Inches y
140	67	127	61
120	62	206	75
191	73	123	62
140	65	137	64
172	71	187	69
190	69	145	65
175	68	197	74
210	75	185	72
132	64	157	66
190	72	194	73
125	61	142	63
192	70	190	71
200	74	142	67
136	63	168	68
162	66	188	70

may be located on the diagram by extending lines from the weight and height scales corresponding to the values 120 and 62 and the dot is located at the point of intersection of the lines. In actual practice it is not necessary to draw the intersecting lines for each dot but they may be located by carefully noting the scale and placing the dots at the proper position. The diagram when completed gives a graphic illustration of the association of these characters, and is useful in presenting a general idea of the relationship. It is not convenient, however, to compare the correlation between these characters by means of this scatter diagram and the correlation between any other pair of characters which may be shown by another scatter diagram, where the class values may be different.

Another scatter diagram where the correlation is not so pronounced is given in Figure 13, page 93. This scatter diagram is made up from the yield records of a number of oat plots. A number of different varieties were grown in ten plots each and the relation between the yield of the first five plots and the yield of the second five plots for each variety is shown in the diagram. Each dot represents one variety and is located as described for Figure 12.

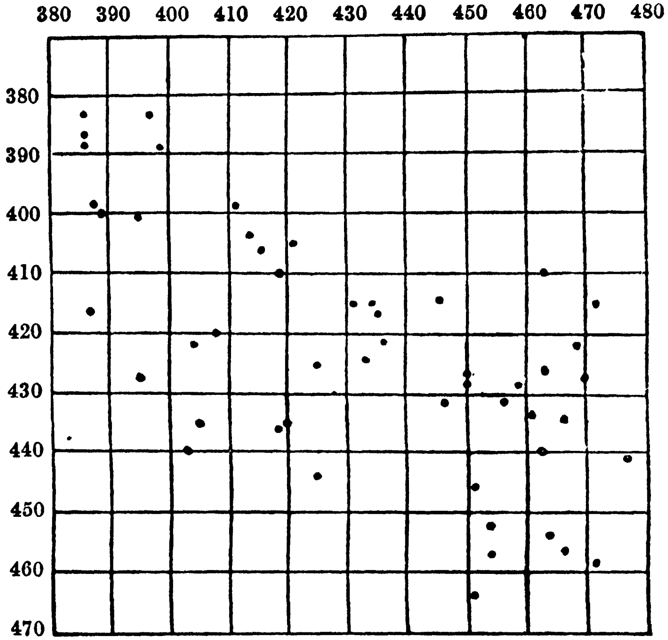


FIG. 13. Scatter diagram showing relation between yields of first five plots and yields of second five plots for a number of varieties of oats.

The illustrations give some idea of how two different lots of data may be compared by means of the scatter diagram, and it is evident that with the data in Figure 12 the relationship is more pronounced than it is with the data in Figure 13. Since it is possible to describe the relationship in both instances by means of qualitative terms only, it becomes evident that what is needed is to be able to express the relationship by means of some definite quantitative value, and such a value to be comparable with other values should be independent of units of measurement.

Relationship may also be shown by means of a graph, as illustrated in Figure 14.

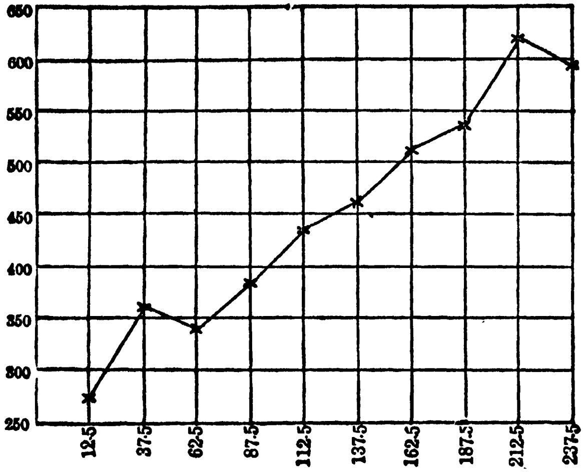


FIG. 14. Graphic illustration of the relation between the yield of cotton at the first picking and the total yield.

The data in the graph are taken from Table 34, which is made up from the data in the first ten classes only of Table 42.

TABLE 34

YIELDS OF SEED COTTON PER PLOT FOR THE FIRST PICKING, AND TOTAL YIELD FOR THE SAME PLOT

MID-POINT VALUE FIRST PICKING	MEAN VALUE TOTAL PICKING
12.5	275.0
37.5	356.0
62.5	344.7
87.5	385.0
112.5	431.0
137.5	462.7
162.5	514.0
187.5	537.5
212.5	623.4
237.5	595.8

In plotting the graph for these data the mid-point values of the yield for the first picking are measured off on the horizontal scale and the mean value of the total picking corresponding to each mid-point value is measured off on the perpendicular scale. For example, to locate the first point on the graph, the mid-point value (on the horizontal scale) is 12.5 and the corresponding mean value for the total picking, 275.0, is measured off on the perpendicular scale and a cross placed where lines extended from the two values would intersect. Each of the other points is located in a similar manner. It should be noted that in order to obtain a better idea of the relationship each scale should be started with the lowest value, increasing the values to the right and upward. It is then possible to associate low values on the horizontal scale with low values on the perpendicular scale.

When such a graph is completed it furnishes some idea of the general tendency as to the amount of association or relation between two characters, but, like the scatter diagram, it does not give any definite numerical value which may be compared with other values obtained from similar data. As will be seen later, a straight line may be fitted to such a graph but even this, while useful for the particular graph, does not furnish a definite means of comparing the data in one graph with those in another unless the data are similar and the same method of grouping has been followed. What is needed, as already stated, is some quantitative expression that will denote the amount of association or correlation. There are several methods for measuring correlation and these will be discussed separately. The most common measure of correlation is the correlation coefficient, which is used to measure linear correlation. By linear correlation is meant that the means as plotted in Figure 14, for example, may be represented by a straight line.

Correlation Coefficient. It has been stated that it is necessary to have some convenient numerical expression to measure relationship. This is especially true since graphic methods do not lend themselves readily to comparisons of one lot of data with another. Then, too, we are usually dealing with a large population and hence there is need for a convenient method for measuring relationship. The correlation coefficient, denoted by r, gives such a numerical expression and may be obtained from the following formula

$$r = \frac{\Sigma xy}{N\,(\sigma_x \sigma_y)}$$

In this formula x represents the deviations of the x values from the mean of x, and y the deviations of the y values from the mean of y, σ_x and σ_y are the standard deviations of the two characters, and N represents the number in the population or the number of pairs of characters.

From this formula it is evident that the value of r depends on the variability of the characters concerned and on the number of pairs of characters that are being studied. The variability of the characters is determined by the standard deviations. In addition to measuring the variability by means of the standard deviations, for the calculation of r it is necessary also to determine the general scatter of the individuals from their respective means. As previously pointed out, it is usually easier to calculate the deviations from an assumed mean rather than from the true mean, and the above formula may therefore be modified to

$$r = \frac{\dfrac{\Sigma D_x D_y}{N} - (c_x c_y)}{\sigma_x \sigma_y}$$

In this formula $\Sigma D_x D_y$, which may be more conveniently called ΣP, represents the product of all deviations of x from the assumed mean of x and all deviations of y from the assumed mean of y, c_x and c_y are the corrections for the mean of x and the mean of y, and σ_x and σ_y are the standard deviations of x and y obtained by working from the assumed mean.

When the number in the population is not too large it may be convenient to determine the relation from the ungrouped material, that is, to obtain the correlation coefficient directly from the items themselves without arranging them in any particular grouping. The method of determining the correlation coefficient from ungrouped data is illustrated in Table 35, using the formula for calculating the correlation coefficient from an assumed mean. The data for x are the average yield of the first five plots and the data for y the average yield of the last five plots of each variety in a test of oats.

TABLE 35

METHOD OF CALCULATING THE CORRELATION COEFFICIENT FROM UNGROUPED MATERIAL. DATA ARE THE AVERAGE YIELD OF THE FIRST FIVE PLOTS, x, AND THE AVERAGE YIELD OF THE LAST FIVE PLOTS, y, OF EACH VARIETY IN A TEST OF OATS

FIRST FIVE PLOTS x	LAST FIVE PLOTS y	D_x	D_y	D_x^2	D_y^2	D_xD_y or ΣP
467	435	32	15	1024	225	480
408	420	−27	0	729	0	0
447	432	12	12	144	144	144
395	429	−40	9	1600	81	−360
475	442	40	22	1600	484	880
399	389	−36	−31	1296	961	1116
436	421	1	1	1	1	1
521	495	86	75	7396	5625	6450
451	447	16	27	256	729	432
461	433	26	13	676	169	338
425	425	−10	5	100	25	− 50
418	410	−17	−10	289	100	170
432	415	− 3	− 5	9	25	15
395	402	−40	−18	1600	324	720
450	427	15	7	225	49	105
421	405	−14	−15	196	225	210
468	423	33	3	1089	9	99
386	388	−49	−32	2401	1024	1568
457	431	22	11	484	121	242
458	428	23	8	529	64	184
		306	208	20)21644	20)10385	13154
N = 20		−236	−111	1082.200000	519.250000	− 410
G_x = 435		20) 70	20) 97	12.250000	23.522500	12744
G_y = 420		c = 3.500	c = 4.850	$\sqrt{1069.950000}$	$\sqrt{495.727500}$	
				σ = 32.710	σ = 22.265	

Correction for Mean of $x = c_x = 3.500$ $\quad \sigma_x = 32.710$

Correction for Mean of $y = c_y = 4.850$ $\quad \sigma_y = 22.265$

$$r = \frac{\frac{\Sigma D_xD_y}{N} - (c_x\, c_y)}{\sigma_x\sigma_y} = \frac{\frac{12744}{20} - (3.500 \times 4.850)}{32.710 \times 22.265} = \frac{637.200000 - 16.975000}{728.288150} = \frac{620.225000}{728.288150} = .852$$

The first step is to assume a mean yield, G, for x and for y, and determine the deviations of the individual items from the assumed mean. These deviations, with their proper signs, are recorded in the columns D_x and D_y. The algebraic sum of each column is obtained and divided by the number of individuals, giving c, or the correction for the mean of x and the mean of y, respectively. The values in columns D_x and D_y are squared and summed and the standard deviation obtained in the usual manner. In addition to the correction values and the standard deviations it is necessary to have the product of the deviation of each item from its respective mean. That is, for the first pair, x deviates from the assumed mean of x by 32 units, and y deviates from the assumed mean of y by 15 units. The product of these deviations is obtained, giving the value 480, and recorded in the column headed D_xD_y, which is more conveniently referred to as ΣP. Such product values are obtained for each pair in turn, being very careful to observe the proper signs. When all the products have been determined their algebraic sum is obtained. This gives the total product deviation of all the pairs of individuals from their respective means. It is necessary to have the average product deviation, and this is obtained by dividing ΣD_xD_y by the number of pairs (in this case 20). Since these products were obtained from an assumed mean it is necessary to correct for this fact, and since we are dealing with products the product of the corrections is therefore obtained and subtracted from the mean product deviation. This gives the correct mean product deviation, 620.225000, as measured from the true mean. The ratio of this value to the product of the standard deviations gives the value of the correlation coefficient, .852. This means that as the yield of x increases there is also an increase in the yield of y.

Frequently it is necessary to obtain the correlation for a large number of individuals, and in such cases, unless the latest types of calculating machines are available, it is usually more convenient to make what is called a double-entry, or correlation, table. Even when calculating machines are available it is helpful to make a correlation table, as such a table gives a graphic illustration of the general relationship of the characters concerned. The method of making a correlation table is illustrated in Figure 15 with the data from Table 35.

						x					
		380.0 389.9	390.0 399.9	400.0 409.9	410.0 419.9	420.0 429.9	430.0 439.9	440.0 449.9	450.0 459.9	460.0 469.9	470.0 479.9	
	380.0–389.9	/	/									2
	390.0–399.9											0
	400.0–409.9		/			/						2
y	410.0–419.9				/		/					2
	420.0–429.9		/	/		/	/		//	/		7
	430.0–439.9							/	/	//		4
	440.0–449.9								/		/	2
		1	3	1	1	2	2	1	4	3	1	19

FIG. 15. Method of making a correlation table. The data are from Table 35.

NOTE: One individual having an x value of 521 and a y value of 495 is omitted to save space in the table.

The first step in making a correlation table is to decide on the number of classes and class interval to be used for each character, as explained in Chapter II on Frequency Distributions. For the data in Table 35 the class interval has been chosen as 10. The classes for each character would be set up on cross-section paper and designated by the name of the characters studied, or by the letters x and y. The letter x is usually taken to represent the columns and the letter y to represent the rows in the table. This gives what is called a double-entry table.

In arranging the classes for a correlation table it is usually better to begin at the same corner of the table with the lowest class for each character, having the classes increase in orderly steps. Tables may be arranged in other ways, but this plan makes it easier to determine at a glance the trend or relationship. For the data in Table 35 each pair of characters would be taken in turn and a mark made in the proper compartment for their numerical values, as shown in Figure 15. For example, the first pair of characters in Table 35 has a value of 467 for x and of 435 for y. The value of x, 467, belongs in column 460.0–469.9 and the value of y, 435, belongs in row 430.0–439.9, and a mark is made in the proper compartment or cell. The second pair is located in column 400.0–409.9 and row

420.0–429.9, and the remaining individuals are similarly located in their proper compartments. The illustration used here gives the location of only a few individuals, but usually we are concerned with a large population and the process would be continued until all the individuals have been recorded in their proper compartments in the table. When all the individuals have been properly located a double-entry, or correlation, table results. For convenience each row or column of such a table may be designated as an *array*.

This method of recording the data in a correlation table does not permit of any convenient means of checking. In order to check the arrangement a second table must be made, and if the two tables do not agree it is necessary to make a third table, or at least a distribution for those classes where the discrepancies occur. For this reason it is much better to make use of the card system, explained in Chapter II on Frequency Distributions. As stated in that chapter, all of the data for an individual will be recorded on one card, and there will be as many cards as there are individuals in the population studied. A table will be set up with the proper classes and the cards will be sorted first for only one of the two characters to be correlated. When this sorting has been completed each pile of cards will be carefully checked to see that none have been placed in the wrong class.

Suppose that this grouping has been done for the y character, then the groups for this character will be kept separate by means of rubber bands or some other convenient arrangement. Each group of cards already sorted for the y character will now be sorted for the x character, and after checking to see that each card is in its proper class the cards in each pile will be counted and the number recorded in the proper compartment of the correlation table. Thus, a frequency distribution is first made for the y character, and then with each lot of cards grouped according to the y character a second frequency distribution is made with each separate group of cards for the x character.

It may be felt that it will take considerable time to copy the data on cards, but, as pointed out earlier, such a system will make a duplicate record of the data and such a duplicate record of important data is always desirable. It might also be pointed out that considerable time will be saved when making correlation tables by using the card system rather than the method of recording pair by pair, as first discussed. The checking is simplified as each pile of cards is checked before it is counted. Another method of checking which is also quickly done is to resort the cards for the x character and then group them for the y character. Often it happens that one character is to be correlated with several others. In such cases it is better to make the first grouping according to the one character that is to be used several times and keep this grouping for all the tables that are to be made.

When the correlation table has been completed the correlation coefficient may be determined as illustrated in Table 36, page 102. The data used are the average weight of kernels per plant in milligrams, x, and the average height of plant in centimeters, y, for oats. The first step is to determine the sum of the frequencies for the x and y characters. The correction for the mean and the standard deviation for each character is obtained by the method already illustrated for determining the correction and standard deviation from grouped material when working from an assumed mean by the unity-step method. The advantage of working from an assumed mean by the unity-step method has already been emphasized, but it is of very great importance in determining the correlation coefficient since by the unity-step method the calculations are much simpler. Also, as will be shown later, for determining the correlation coefficient it is not necessary to multiply c or the standard deviation by the class interval.

Having determined the corrections and the standard deviations it remains only to obtain the values for the product column. As stated in connection with the method for determining r from ungrouped data, it is necessary to determine the deviation of each individual from its respective mean, M_x or M_y, and obtain the product of these deviations. Since there may be a number of individuals in any one class these product values may be obtained by groups or classes rather than by taking each individual separately. As it is necessary to determine the deviations of the x characters from the mean of x and of the y characters from the mean of y,

TABLE 36

METHOD OF CALCULATING THE CORRELATION COEFFICIENT FROM GROUPED MATERIAL. DATA ARE THE AVERAGE WEIGHT OF KERNELS PER PLANT IN MILLIGRAMS, x, AND THE AVERAGE HEIGHT OF PLANT IN CENTIMETERS, y, FOR OATS

		x								f	D_y	fD_y	fD^2_y	D_{xa}	$D_{xa}D_y$ or ΣP
		12	13	14	15	16	17	18	19						
	55.0–59.9				2					2	−3	− 6	18	2	−6
	60.0–64.9		2	5	3	1				11	−2	−22	44	3	−6
	65.0–69.9	1	12	22	8	2				45	−1	−45	45	−2	2
y	70.0–74.9	3	27	52	43	10	2	2	1	140	0			49	0
	75.0–79.9		11	50	42	15	3		1	122	1	122	122	75	75
	80.0–84.9	1	2	25	28	14	2	1		73	2	146	292	62	124
	85.0–89.9			4	2		1			7	3	21	63	5	15
												289	584	196	216
f		5	54	158	128	42	8	3	2	400		−73		−2	−12
												216		194	204
D_x		−2	−1	0	1	2	3	4	5						
fD_x		−10	−54		128	84	24	12	10	= 194					
fD^2_x		20	54		128	168	72	48	50	= 540					

$$c_x = \frac{194}{400} = .485$$

$$\sigma_x = \sqrt{\frac{540}{400} - (.485)^2} = \sqrt{1.350000 - .235225} = 1.056$$

$$c_y = \frac{216}{400} = .540$$

$$\sigma_y = \sqrt{\frac{584}{400} - (.540)^2} = \sqrt{1.460000 - .291600} = 1.081$$

$$r = \frac{\frac{\Sigma P}{N} - (c_x\, c_y)}{\sigma_x\, \sigma_y}$$

$$r = \frac{\frac{204}{400} - (.485 \times .540)}{1.056 \times 1.081}$$

$$= \frac{.510000 - .261900}{1.141536}$$

$$= \frac{.248100}{1.141536} = .217$$

it is more convenient to obtain the values in the product column (ΣP) in two steps. The values in the column D_{xa} are first determined. These values represent the sum of the deviations in each y array measured from the assumed mean of x. For example, for y class 55.0–59.9 there are 2 individuals in x class 15. It is noted that these 2 individuals, as well as all individuals in class 15, deviate by $+1$ from the assumed mean of x. We multiply these 2 individuals by $+1$, giving the product $+2$ for column D_{xa}. For the next y class, 60.0–64.9, there are 2 individuals in class 13 and these deviate by -1 from the assumed mean, so the product is 2×-1, or -2. In class 14 there are 5 individuals, but since this is the class of the assumed mean the deviation of this class from the assumed mean is 0, and the product is 5×0, or 0. In class 15 there are 3 individuals that deviate by $+1$ from the assumed mean, and the product is $3\times+1$, or $+3$. In class 16 there is 1 individual which deviates by $+2$ from the assumed mean, and the product is $1\times+2$, or $+2$. The sum of these products, -2, $0, +3, +2$, equals $+3$. Care should be taken to observe the signs in obtaining the sum. Proceeding in the same manner for the other arrays the results given in column D_{xa} are obtained. The complete steps in determining these values are given here.

y Class	Values in y Classes Multiplied by Their Respective Deviations from the Mean of x	Sum or D_{xa}
55.0–59.9	$2\times+1$	$+2$
60.0–64.9	$2\times-1,\ 5\times0,\ 3\times+1,\ 1\times+2$	$+3$
65.0–69.9	$1\times-2,\ 12\times-1,\ 22\times0,\ 8\times+1,\ 2\times+2$	-2
70.0–74.9	$3\times-2,\ 27\times-1,\ 52\times0,\ 43\times+1,\ 10\times+2,\ 2\times+3,\ 2\times+4,\ 1\times+5$	$+49$
75.0–79.9	$11\times-1,\ 50\times0,\ 42\times+1,\ 15\times+2,\ 3\times+3,\ 1\times+5$	$+75$
80.0–84.9	$1\times-2,\ 2\times-1,\ 25\times0,\ 28\times+1,\ 14\times+2,\ 2\times+3,\ 1\times+4$	$+62$
85.0–89.9	$4\times0,\ 2\times+1,\ 1\times+3$	$+5$
		$+196$ -2 $+194$

When all the products have been obtained for column D_{xa} they are summed, having regard to the signs, and since the deviations have been obtained from the assumed mean of x and include all of the individuals in the table, the sum of these products should be the same as the sum of fD_x, which deviations have also been obtained from the same assumed mean of x, with all the individuals in the table represented. If the sum of D_{xa} equals the sum of fD_x it shows that the process of obtaining the products of the various arrays from the x mean has been correctly carried out.

Now, to obtain the total product $D_{xa}D_y$, or ΣP, it is necessary only to multiply the values in column D_{xa} by the corresponding values in column D_y. Thus, for the first result $D_{xa} = +2$ and $D_y = -3$, and the product is -6. Multiplying the remaining values in D_{xa} by their corresponding values in D_y gives the values in the column ΣP. These values are summed, having regard to signs, and this sum is the total product deviation of all the individuals from the assumed means. This sum, divided by the number of individuals or pairs, gives the mean product deviation. Since this has been obtained from an assumed mean it is necessary to correct for this fact, and since we are dealing with products the product of the corrections is obtained and subtracted from the mean product. The remainder is divided by the product of the standard deviations, giving the value for r. Substituting the different values in the formula the correlation coefficient for these two characters is

$$r = \frac{\frac{\Sigma P}{N} - (c_x c_y)}{\sigma_x \sigma_y}$$

$$r = \frac{\frac{204}{400} - (.485 \times .540)}{1.056 \times 1.081} = .217$$

It is important to note that when using the corrections and the standard deviations in obtaining the correlation coefficient we are dealing with values obtained by the unity-step method and yet do not multiply them by the class interval. This is correct since the values in the product column, ΣP, have also been obtained by the unity-step method and therefore all values must be treated alike. Great care should be exercised to see that this is done in all cases, since if one value is multiplied by the class interval it is necessary that all values be multiplied by the class interval, and errors may arise if this fact is not kept in mind. For this reason it is better to determine the value of the correlation coefficient before considering the class interval. If the means and standard deviations are desired the corrections and standard deviations may then be multiplied by the proper class interval.

The correlation coefficient (.217) as just determined gives the measure of relationship between the two characters in question, that is, as the average height of plant increases there is some increase in the average weight of kernels per plant. The closeness of this relationship may be understood better when it is known what the highest value of a correlation coefficient may be. In other words, if there should be perfect correlation what will be the numerical value of r. This may be shown by the following illustration.

Let us now assume the simplest case of relationship. Suppose we have some simple characters which may be expressed in units and the unit of the first character, which we shall call x, is assumed to be represented by the value 1, and similarily the measurement of the second character, y, that is related to the first is represented by the value 1. Then assuming we have a second individual in which the first character has increased by 1 unit and likewise the second character of the same individual has increased by 1 unit, we follow this same scheme for the 3rd, 4th, and succeeding individuals up to 10. Then we would have the following values for x and y:

x	y
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10

In order to show the relationship the ten individuals are placed in a correlation table.

TABLE 37

ILLUSTRATING A CASE OF PERFECT POSITIVE CORRELATION

		x										f	D_y	fD_y	fD^2_y	D_{xa}	$D_{xa}D_y$ or ΣP
		1	2	3	4	5	6	7	8	9	10						
y	1	1										1	−4	−4	16	−4	16
	2		1									1	−3	−3	9	−3	9
	3			1								1	−2	−2	4	−2	4
	4				1							1	−1	−1	1	−1	1
	5					1						1	0				
	6						1					1	1	1	1	1	1
	7							1				1	2	2	4	2	4
	8								1			1	3	3	9	3	9
	9									1		1	4	4	16	4	16
	10										1	1	5	5	25	5	25
														15	85	15	85
f		1	1	1	1	1	1	1	1	1	1	10		−10		−10	
D_x		−4	−3	−2	−1	0	1	2	3	4	5			5		5	
fD_x		−4	−3	−2	−1		1	2	3	4	5	= 5					
fD^2_x		16	9	4	1		1	4	9	16	25	= 85					

$$c_x = \frac{5}{10} = .5$$

$$\sigma_x = \sqrt{\frac{85}{10} - (.5)^2} = \sqrt{8.50 - .25} = \sqrt{8.25}$$

$$c_y = \frac{5}{10} = .5$$

$$\sigma_y = \sqrt{\frac{85}{10} - (.5)^2} = \sqrt{8.50 - .25} = \sqrt{8.25}$$

The formula is

$$r = \frac{\frac{\Sigma P}{N} - (c_x c_y)}{\sigma_x \sigma_y}$$

Substituting the values obtained

$$r = \frac{\frac{85}{10} - (.5 \times .5)}{\sqrt{8.25} \times \sqrt{8.25}}$$

$$= \frac{8.50 - .25}{8.25}$$

$$= \frac{8.25}{8.25} = 1.00$$

The first individual which has the x character represented by 1 also has the y character represented by 1, and it belongs in the first compartment in the upper left-hand corner of the table. The second individual in which the x character is represented by 2 units also has the y character represented by 2 units, and it belongs in the compartment one unit below and one unit to the right of the first compartment, and so on for all of the ten individuals.

Using the formula for the correlation coefficient by working from an assumed mean we may now proceed to determine r. Assuming the mean for each character to be at 5, we obtain the various columns in the usual way. The corrections for x and y are the same, .5, and the standard deviation for each character is likewise the same, $\sqrt{8.25}$. Substituting these values in the formula

$$r=\frac{\frac{\Sigma P}{N}-(c_xc_y)}{\sigma_x\sigma_y}$$

$$r=\frac{\frac{85}{10}-(.5\times.5)}{\sqrt{8.25}\times\sqrt{8.25}}=\frac{8.50-.25}{8.25}=1.00$$

Since we have assumed a case of perfect relationship where the same increase in the x character is found in the y character, this shows that the numerical value of r can never be more than 1.00. For positive correlation the value of r may vary from 0 to 1.00.

Now, if another case is assumed in which when the y character is represented by 1 unit the x character is represented by 10 units, and when the y character is represented by 2 units the x character is represented by 9 units, and so on until the y character is represented by 10 units and the x character by 1 unit, the columns for x and y will be as given on page 109.

TABLE 38

ILLUSTRATING A CASE OF PERFECT NEGATIVE CORRELATION

		x										*f*	D_y	fD_y	fD^2_y	D_{xa}	$D_{xa}D_y$ or ΣP
		1	2	3	4	5	6	7	8	9	10						
y	1										1	1	−4	−4	16	4	−16
	2									1		1	−3	−3	9	3	− 9
	3								1			1	−2	−2	4	2	− 4
	4							1				1	−1	−1	1	1	− 1
	5						1					1	0				
	6					1						1	1	1	1	−1	− 1
	7				1							1	2	2	4	−2	− 4
	8			1								1	3	3	9	−3	− 9
	9		1									1	4	4	16	−4	−16
	10	1										1	5	5	25	−5	−25
														15	85	10	−85
f		1	1	1	1	1	1	1	1	1	1	10		−10		−15	
D_x		−5	−4	−3	−2	−1	0	1	2	3	4			5		− 5	
fD_x		−5	−4	−3	−2	−1		1	2	3	4	= −5					
fD^2_x		25	16	9	4	1		1	4	9	16	= 85					

$$c_x = \frac{-5}{10} = -.5$$

$$\sigma_x = \sqrt{\frac{85}{10} - (-.5)^2} = \sqrt{8.50 - .25} = \sqrt{8.25}$$

$$c_y = \frac{5}{10} = .5$$

$$\sigma_y = \sqrt{\frac{85}{10} - (.5)^2} = \sqrt{8.50 - .25} = \sqrt{8.25}$$

The formula is

$$r = \frac{\frac{\Sigma P}{N} - (c_x c_y)}{\sigma_x \sigma_y}$$

Substituting the values obtained

$$r = \frac{\frac{-85}{10} - (-.5 \times .5)}{\sqrt{8.25} \times \sqrt{8.25}}$$

$$= \frac{-8.50 - (-.25)}{8.25}$$

$$= \frac{-8.25}{8.25} = -1.00$$

x	y
10	1
9	2
8	3
7	4
6	5
5	6
4	7
3	8
2	9
1	10

Placing these values in the proper compartments the results are as given in Table 38, page 108.

Using the same formula for r as before and substituting the values from Table 38 in this formula we have

$$r=\frac{\frac{\Sigma P}{N}-(c_x c_y)}{\sigma_x \sigma_y}$$

$$r=\frac{\frac{-85}{10}-(-.5\times .5)}{\sqrt{8.25}\times\sqrt{8.25}}=\frac{-8.50-(-.25)}{8.25}=-1.00$$

The correlation coefficient is -1.00, due to the fact that ΣP is minus. This shows that it may be possible to have a negative value for r, but that it will never be more numerically than -1.00. Thus it is clear that the limits of r are from +1.00 to -1.00 in value, or the variation in r may be from 0 to +1.00 or -1.00. A positive value for r indicates that as the value of one character increases there is also an increase in the value of the second character. When a negative value for r is obtained it means that as the value for one character increases the value of the second character decreases.

Another illustration of a correlation table and the steps for calculating the correlation coefficient are shown in Table 39, page 110, showing the correlation between the weight of grain after threshing, x, and the total weight of heads, y, of kaoliang plants.

TABLE 39

CORRELATION BETWEEN WEIGHT OF GRAIN AFTER THRESHING, x, AND TOTAL WEIGHT OF HEADS, y, OF KAOLIANG PLANTS

	x								f	D_y	fD_y	fD^2_y	D_{xa}	$D_{xa}D_y$ or ΣP
	2.00 9.99	10.00 17.99	18.00 25.99	26.00 33.99	34.00 41.99	42.00 49.99	50.00 57.99	58.00 65.99						
y 3.00–10.99	130								130	-2	-260	520	-260	520
11.00–18.99	70	289							3 9	-1	-359	359	-429	429
19.00–26.99		178	138						316	0			-178	0
27.00–34.99			168	41					209	1	209	209	41	41
35.00–42.99				84	3				87	2	174	348	90	180
43.00–50.99				3	37	2			42	3	126	378	83	249
51.00–58.99					7	15			22	4	88	352	59	236
59.00–66.99						1	1		2	5	10	50	7	35
67.00–74.99							2	2	4	6	24	144	18	108
75.00–82.99							1		1	7	7	49	4	28
f	200	467	306	128	47	18	4	2	1172		638 -619 19	2409	302 -867 -565	1826
D_x	-2	-1	0	1	2	3	4	5						
fD_x	-400	-467		128	94	54	16	10 =	-565					
fD^2_x	800	467		128	188	162	64	50 =	1859					

$$c_x = \frac{-565}{1172} = -.482$$

$$\sigma_x = \sqrt{\frac{1859}{1172} - (-.482)^2} = \sqrt{1.586177 - .232324} = 1.164$$

$$c_y = \frac{19}{1172} = .016$$

$$\sigma_y = \sqrt{\frac{2409}{1172} - (.016)^2} = \sqrt{2.055461 - .000256} = 1.434$$

$$r = \frac{\frac{1826}{1172} - (-.482 \times .016)}{1.164 \times 1.434}$$

$$= \frac{1.558020 - (-).007712}{1.669176}$$

$$= \frac{1.565732}{1.669176} = .938$$

The various steps in the calculation are followed through as in the previous illustrations. Substituting the different values in the formula, $r=.938$. This is a much higher value for r than was obtained for the data in Table 36. Since r can never be higher than 1.00, this correlation coefficient of .938 shows a very high relationship or close association for the two characters concerned. More will be said in later chapters regarding the evaluation or interpretation of correlation coefficients.

It is important to note in Table 39 that the correction for y is positive while the correction for x is negative. Therefore, in obtaining the product of the corrections it is necessary to multiply a positive by a negative value and the product will be negative. When this negative quantity is subtracted from the mean product deviations it is actually added. It is important to stress here that at this particular point in the calculations there is danger of errors occurring unless the products of the corrections are handled very carefully according to their signs. It is possible to have the following signs for the corrections and for their products:

$$
\begin{gathered}
+\times+=+\\
-\times-=+\\
-\times+=-\\
+\times-=-
\end{gathered}
$$

In the first two cases, where the product is positive and there is a positive value for the mean product deviation, the product of the corrections is subtracted from the mean product deviation. In the other cases, where the product is negative, if the mean product deviation is positive then the negative quantity is subtracted, which in reality means that it is added. When the mean product deviation is negative then the opposite is true. Thus, in subtracting a positive correction product from a mean product deviation that is negative we actually add, and with a negative correction we subtract.

Another example of correlation is shown in Table 40, page 112. The data are the number of kernels per plant, x, and the average weight of kernels per plant in milligrams, y, for oats.

TABLE 40

CORRELATION BETWEEN NUMBER OF KERNELS PER PLANT, x, AND THE AVERAGE WEIGHT OF KERNELS PER PLANT IN MILLIGRAMS, y, FOR OATS

	x										f	D_y	fD_y	fD^2_y	D_{xa}	$D_{xa}D_y$ or ΣP
	0.0 49.9	50.0 99.9	100.0 149.9	150.0 199.9	200.0 249.9	250.0 299.9	300.0 349.9	350.0 399.9	400.0 449.9	450.0 499.9						
11.50–12.49			1								1	–5	– 5	25	– 2	10
12.50–13.49		1	2			1					4	–4	– 16	64	– 6	24
13.50–14.49	1	1	1	5	1	4	1	1	1		16	–3	– 48	144	– 1	3
14.50–15.49		2	7	9	6	15	7	2	1	1	50	–2	–100	200	15	–30
15.50–16.49		2	16	12	20	11	7	3	1		72	–1	– 72	72	–12	12
16.50–17.49		5	22	24	25	11	14	1	1		103	0			–37	0
y 17.50–18.49	1	7	21	22	20	9	3				83	1	83	83	–74	–74
18.50–19.49		1	12	11	5	6	1			1	37	2	74	148	–25	–50
19.50–20.49		2	6	6	3	1		1			19	3	57	171	–20	–60
20.50–21.49			4	5	1						10	4	40	160	–13	–52
21.50–22.49			3	1							4	5	20	100	– 7	–35
22.50–23.49											0	6	0	0	0	0
23.50–24.49				1							1	7	7	49	– 1	– 7
													281	1216	15	–308
f	2	21	95	96	81	58	33	8	4	2	400		–241		–198	49
													40		–183	–259
D_x	–4	–3	–2	–1	0	1	2	3	4	5						
fD_x	–8	–63	–190	–96		58	66	24	16	10=	–183					
fD^2_x	32	189	380	96		58	132	72	64	50=	1073					

$$c_x = \frac{-183}{400} = -.457$$

$$\sigma_x = \sqrt{\frac{1073}{400} - (-.457)^2} = \sqrt{2.682500 - .208849} = 1.573$$

$$c_y = \frac{40}{400} = .100$$

$$\sigma_y = \sqrt{\frac{1216}{400} - (.100)^2} = \sqrt{3.040000 - .010000} = 1.741$$

$$r = \frac{\frac{-259}{400} - (-.457 \times .100)}{1.573 \times 1.741}$$

$$= \frac{-.647500 - (-).045700}{2.738593}$$

$$= \frac{-.601800}{2.738593} = -.220$$

NOTE. These data on the number of kernels have been obtained by counting, and naturally there would be no fractional numbers. However, as a guide in the arrangement of classes the class limits are written so as to indicate definitely what is included in each class. For example, the first class, 0.0-49.9, indicates that this class includes all numbers below 50, and class 50.0-99.9 includes 50 and all numbers below 100.

The various steps for calculating the correlation coefficient are also shown in this table. It is to be noted that the summation for the column $D_{xa}D_y$ is negative, and that the product of the corrections is negative. In this case it is necessary to subtract a negative quantity from a negative quantity, which means that the difference is taken since in subtracting a negative quantity it becomes plus. Since there is a negative quantity for the ΣP column and therefore a negative quantity to be divided by the product of the standard deviations, the sign of the correlation is negative. This means for this particular case that as the number of kernels per plant increases there is a tendency for the average weight of kernels to decrease.

It has already been stated that when using the unity-step method in determining the correlation coefficient it is not necessary to correct for the class interval. To prove this statement Table 41 is given with the classes for x and y omitted. The calculations are carried out in the usual manner and by substituting the values obtained in the formula for r the correlation coefficient is found to be .665.

In a correlation table the class limits or the arrangement of the classes is necessary only for determining the location of the individuals in the table. After the table has been completed it is not necessary to refer again to the arrangement of classes with regard to the size or class interval, and Table 41 given on page 114 is merely to emphasize this point.

Another example of the use of correlation is given in Table 42, page 115, showing the correlation between the total yield of cotton, x, and the weight of cotton at the first picking, y, for the same plot. The value of r is .644, which indicates a fairly close relation between the first picking and total yield per plot.

In calculating the correlation coefficient it is evident that r is really a ratio value. This may be made clearer by referring to Figure 16, page 116, in which the results from Table 37 are used.

TABLE 41

CORRELATION COEFFICIENT DETERMINED WITHOUT REFERENCE TO CLASS INTERVAL. DATA ARE AVERAGE NUMBER OF SPIKELETS PER CULM, x, AND TOTAL YIELD OF PLANT IN GRAMS, y, FOR OATS

	x								f	D_y	fD_y	fD^2_y	D_{xa}	$D_{xa}D_y$ or ΣP
y	3								3	−2	−6	12	−9	18
	4	13	14	11	6	1		1	50	−1	−50	50	−40	40
	2	7	30	40	19	6	2		106	0			−13	0
		1	13	31	45	16	3		109	1	109	109	71	71
			2	15	33	20	9	1	80	2	160	320	102	204
					9	18	11	4	42	3	126	378	94	282
						4	2	1	7	4	28	112	18	72
							1	1	2	5	10	50	7	35
						1			1	6	6	36	2	12
											439	1067	294	734
f	9	21	59	97	112	66	28	8	400		−56		−62	
D_x	−3	−2	−1	0	1	2	3	4			383		232	
fD_x	−27	−42	−59		112	132	84	32	= 232					
fD^2_x	81	84	59		112	264	252	128	= 980					

$$c_x = \frac{232}{400} = .580 \qquad c_y = \frac{383}{400} = .957$$

$$\sigma_x = \sqrt{\frac{980}{400} - (.580)^2} = \sqrt{2.450000 - .336400} = 1.454 \qquad \sigma_y = \sqrt{\frac{1067}{400} - (.957)^2} = \sqrt{2.667500 - .915849} = 1.323$$

The values are obtained without reference to classes by the unity-step method, and are substituted directly in the formula for r. Therefore

$$= \frac{\frac{734}{400} - (.580 \times .957)}{1.454 \times 1.323} = \frac{1.835000 - .555060}{1.923642} = \frac{1.279940}{1.923642} = .665$$

TABLE 42

CORRELATION BETWEEN THE TOTAL YIELD OF COTTON, x, AND THE WEIGHT OF COTTON AT THE FIRST PICKING, y

	x																
	100.0 149.9	150.0 199.9	200.0 249.9	250.0 299.9	300.0 349.9	350.0 399.9	400.0 449.9	450.0 499.9	500.0 549.9	550.0 599.9	600.0 649.9	650.0 699.9	700.0 749.9	750.0 799.9	800.0 849.9	850.0 899.9	f
0.0– 24.9	1	1		3		1	1										7
25.0– 49.9		1	2	5	3	3	3	2		2							21
50.0– 74.9		1	6	7	8	1	4	3	1		1	1					33
75.0– 99.9			3	11	2	19	10	4	3	1	1			1			55
100.0–124.9			1	7	10	14	12	10	11	5	4	1					75
125.0–149.9			1	1	6	14	14	13	11	5	3	4	1				73
150.0–174.9				2		7	11	11	9	4	7	3	3		2		59
175.0–199.9					2	3	2	7	5	13	3	3	1		1		40
y 200.0–224.9								5	3	5	5	6	4	2		1	31
225.0–249.9							1	2	4	8	4	1	1	3			24
250.0–274.9								1	6	1	1	4	1				14
275.0–299.9								1	1	1			1				4
300.0–324.9									2		3						5
325.0–349.9												1			1	1	3
350.0–374.9														1			1
375.0–399.9												2	1				3
400.0–424.9														2			2
f	1	3	13	36	31	62	58	59	56	45	32	26	13	9	4	2	450

Values obtained on the unity-step basis

$c_x = 1.000$ $c_y = -.691$

$\sigma_x = 2.811$ $r = .644$ $\sigma_y = 2.787$

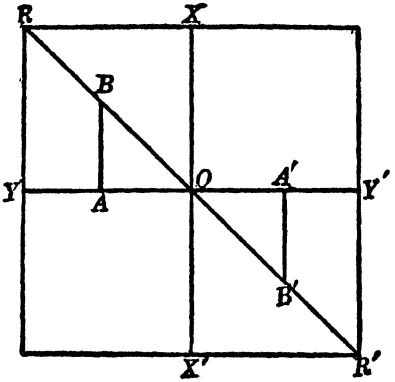


FIG. 16.

A line RR' has been drawn through the means of the rows, cutting through the point of intersection of lines XX' and YY', which represent the means of x and y, respectively. If a point A is located on line YY' and a perpendicular erected to intersect line RR' at B, we have the triangle OAB, in which $OA = AB$. Since OA and AB are distances measured below the mean of x and the mean of y, the sign of each is minus. If a point A' is located on line YY' and a perpendicular let fall to intersect line RR' at B', the signs of OA' and $A'B'$ are plus. Since the line RR' is drawn through the means of the rows, then the ratio of OA to AB may be taken to represent the relationship between the two characters, x and y. In this instance since $OA = AB$ the ratio OA/AB is 1.00, which is true for perfect positive correlation. Now OA may represent any value of x and AB may represent any value of y. Therefore we may substitute x and y for these values and denote a general relationship of x to y. It is also possible to have the ratio of y to x. If the values in Table 38 were used and the same reasoning followed it is evident that the same numerical value for the ratio would be found, but the sign would be negative.

This ratio value shows the relation between the two characters, and this is true for any correlation table. It should be kept in mind, however, that the ratio values are determined on the basis of the scale of measurement used, and in order to obtain a value independent of the scale of measurement it is desirable to express this ratio in terms of some measure common to both x and y. This will be made clear by reference to Figure 17, where the means of the rows for the data in Table 39 are represented by dots.

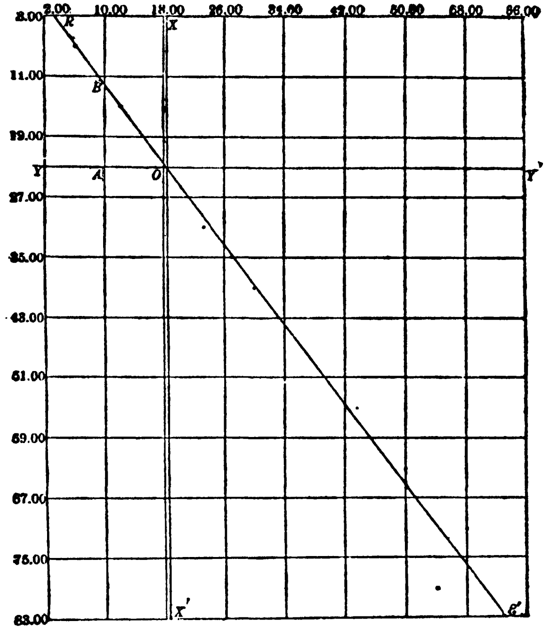


FIG. 17. Relation between the weight of grain after threshing and the total weight of heads of kaoliang plants.

A line RR' has been drawn by inspection to fit these means. Details for fitting the line by more accurate methods will be given later. A point A has been located on the line YY' at x value 10. From this point a perpendicular has been drawn to intersect RR' at B. Since point A is located at x 10, the distance of this point from the mean of x (18.144) is 8.144. By measuring the length of line AB using the same scale of measurement it is found to equal 10.880 units. Dividing OA by AB we have

$$8.144/10.880 = .749$$

This is the ratio denoting the relationship between x and y, and we may substitute x and y for OA and AB to give a general expression of relationship. It must be clear that for any x and y values the ratio will be the same, .749.

If the variation as expressed by the standard deviation were the same for x and y, this ratio would express very closely the value of the correlation coefficient. Since the variation is not the same as measured by the standard deviation it is better to consider the standard deviation before obtaining a ratio which will approximate r. This is done by dividing OA by the standard deviation of x and dividing AB by the standard deviation of y. Thus we have

$$\frac{OA/\sigma_x}{AB/\sigma_y}$$

Substituting the values for the data in Table 39

$$\frac{8.144/9.312}{10.880/11.472}=\frac{.875}{.948}=.923$$

This value approaches very closely the value for the correlation coefficient, .938. It is evident that the correlation coefficient may be represented by a ratio in the manner just described.

Regression Lines. The determination of the correlation coefficient is of value in judging the relationship of data that have already been studied. Another important use of correlation is to predict on the basis of past experience or results already obtained what may be expected to occur with other similar material. For example, if it is found that rainfall at a certain period of the year is related to the yield of wheat for any given locality, then it is important to be able to predict on the basis of the amount of rainfall what the yield for a particular season may be. In other words, for a particular amount of rainfall what may be the expected yield? Again, if the weight and height of men are correlated to a certain degree, what will be the expected weight of a group of men of a given height? The analysis of correlation is useful in answering these and similar questions by means of prediction lines or regression lines. In Figure 17 the line was drawn by inspection, as stated, but for purposes of prediction the lines should be fitted by more exact methods. The methods for fitting these prediction or regression lines will now be discussed.

Let us assume the simplest case for the line denoting the regression, in which the means of the rows fall exactly on the straight line, as for example *RR* in Figure 18. (*Adapted from Yule.*)

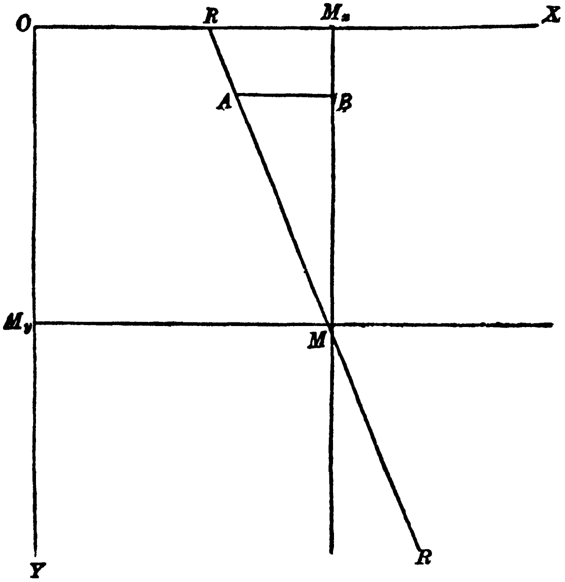


Fig. 18.

Let the slope of this line *RR* to the vertical line MM_x, that is the tangent of angle *BMA*, or the ratio *AB* to *BM*, be represented by b_1, and let the deviations of the x values be measured from the mean of x (M_x) and the deviations of y be measured from the mean of y (M_y). Now, for any one row of y values in which we have n observations, we have

$$\Sigma(x) = nb_1y$$

Then, since this gives the values for any one row we would have for the whole table

$$\Sigma(x) = b_1\Sigma(ny)$$

Since y represents deviations from the mean of y and since for the whole table there will be as many negative deviations as positive deviations, therefore

$$\Sigma(ny)=0$$

and

$$\Sigma(x)=b_1\Sigma(ny)=0$$

Since x represents the deviations from the mean of x and since Σx equals 0, then the line BM must pass through the mean of x and therefore M_x is the mean of x.

In a similar manner it will later be shown that another line that passes through the means of the columns will also pass through this same point M. From this it follows that the regression lines will intersect at M, which is the mean of the whole distribution, or, in other words, the regression lines will intersect at the same point that the line representing the mean of x and the line representing the mean of y cross each other.

We may now proceed to determine the value of b_1 by working from the mean product deviation, as already found from the product columns of our correlation tables. This mean product deviation equals $\Sigma xy/N$ and may be represented by p.
Hence

$$p=\Sigma xy/N$$

Now, for the product deviations of any one row of the y distribution, using the equation representing the deviations of any one row

$$\Sigma(x)=n\ b_1y$$

multiplying by y, then

$$\Sigma(xy)=n\ b_1y^2$$

This represents the product values for any one row. For the whole table the product values are represented by

$$\Sigma(xy)=b_1\ \Sigma(ny^2)$$

Since

$$\Sigma(ny^2)=N\sigma^2y$$

we have

$$\Sigma(xy) = N\, b_1\, \sigma^2_y$$

Since

$$p = \Sigma xy/N$$

we have by substitution and transposition

$$b_1\, \sigma^2_y = \Sigma xy/N = p$$

and

$$b_1 = p/\sigma^2_y$$

From the formula for the correlation coefficient

$$r = \frac{\frac{\Sigma xy}{N}}{\sigma_x \sigma_y} \quad \text{or} \quad r = \frac{p}{\sigma_x \sigma_y}$$

we have

$$p = r\sigma_x\sigma_y$$

Substituting this in the formula

$$b_1 = p/\,\sigma^2_y$$

we have

$$b_1 = \frac{r\sigma_x\sigma_y}{\sigma^2_y}$$

Therefore by cancellation

$$b_1 = r\frac{\sigma_x}{\sigma_y}$$

Now, since b_1 is the ratio of AB to BM, and since AB is measured on the x scale and BM on the y scale, it is clear that the ratio of any distance measured vertically from the line MM_x to the line RR to any distance marked off on BM will give the same ratio as AB to BM. Therefore, since AB may represent any deviation of x and BM any deviation of y, then b_1 may be equal to x/y. Substituting this value x/y in the formula above we have

$$\frac{x}{y} = r\frac{\sigma_x}{\sigma_y}$$

Therefore

$$x = r\frac{\sigma_x}{\sigma_y}y$$

Now x represents deviations from the mean of x, and y deviations from the mean of y. For the location, or plotting, of our line we need to express them in the original units or values. That is, we need to consider the original scales of measurement. To do this we make $x=(X-M_x)$ and $y=(Y-M_y)$, and the formula becomes by substitution

$$X-M_x=r\frac{\sigma_x}{\sigma_y}(Y-M_y)$$

In this formula M_x and M_y represent the means of x and y, respectively, σ_x and σ_y are the standard deviations of x and y, respectively, and r is the correlation coefficient. This gives the regression equation for x on y.

It is possible to follow through the same argument for the means of columns, and the regression equation in this instance becomes

$$y=r\frac{\sigma_y}{\sigma_x}x$$

Hence

$$Y-M_y=r\frac{\sigma_y}{\sigma_x}(X-M_x)$$

These equations may now be applied to the values obtained in Table 36. From the correction values, M_x and M_y may be obtained in the usual manner, giving $M_x=14.485$ and $M_y=75.200$. The standard deviations are multiplied by the class interval, giving the true standard deviation, or

$$\sigma_x=1.056\times 1=1.056$$
$$\sigma_y=1.081\times 5=5.405$$

It should be noted that for the determination of regression the standard deviation *must always be multiplied by the proper class* *interval.* Substituting the values for the means, standard deviations, and r in the formula for the regression of x on y, we have

$$X-14.485=.217\frac{1.056}{5.405}(Y-75.200)$$
$$X-14.485=.217\times.195\ (Y-75.200)$$
$$X-14.485=.042\ (Y-75.200)$$
$$X-14.485=.042Y-3.158$$
$$X=.042Y+11.327$$

In order to fit these lines to the correlation table it is necessary to select certain values of y and determine the calculated values of x from the equation

$$X=.042Y+11.327$$

For $Y=60$
$$X=.042(60)+11.327$$
$$X=2.520+11.327$$
$$X=13.847$$

For $Y=85$
$$X=.042(85)+11.327$$
$$X=3.570+11.327$$
$$X=14.897$$

For $Y=75.200$
$$X=.042(75.200)+11.327$$
$$X=3.158+11.327$$
$$X=14.485$$

Connecting these points with a straight line and extending the line to the limits of the correlation table we have a graph of the regression line. (See Figure 19, on page 124.)

For the regression of y on x for the data in Table 36 we have the equation

$$Y-75.200=.217\frac{5.405}{1.056}(X-14.485)$$
$$Y-75.200=.217\times5.118\ (X-14.485)$$
$$Y-75.200=1.111\ (X-14.485)$$
$$Y-75.200=1.111X-16.093$$
$$Y=1.111X+59.107$$

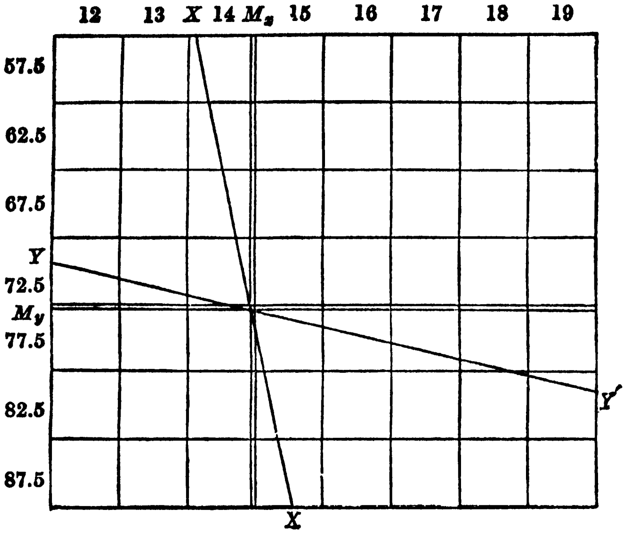


FIG. 19. Regression lines fitted to the data in Table 36. The line XX' shows the regression of x on y and the line YY' shows the regression of y on x

It should be noted that the value of Y cannot be calculated from the equation for x on y. Selecting certain values of x and determining the calculated values of y from the equation

$$Y=1.111X+59.107$$

For $X=13$
$$Y=1.111(13)+59.107$$
$$Y=14.443+59.107$$
$$Y=73.550$$

For $X=18$
$$Y=1.111(18)+59.107$$
$$Y=19.998+59.107$$
$$Y=79.105$$

For $X=14.485$
$$Y=1.111\ (14.485)+59.107$$
$$Y=16.093+59.107$$
$$Y=75.200$$

It is noted that when Y is taken as 75.200 (the mean of y) X is found to be 14.485 (the mean of x), and likewise when the mean of x, 14.485, is taken for X, Y is found to be 75.200. This shows that the regression lines pass through the point of intersection of the line for the mean of x and the line for the mean of y.

Connecting the three points for the Y values by means of a straight line we have the regression line of y on x, as illustrated in Figure 19. Such lines are prediction lines. This may be illustrated by using the regression of x on y and predicting the expected value of x for y class 80.0–84.9. Substituting the mid-point value of this class in the equation we have

$$\text{For } Y=82.5 \qquad \begin{aligned} X&=.042(82.5)+11.327\\ X&=3.465+11.327\\ X&=14.792 \end{aligned}$$

This value, 14.792, is the predicted value of x. The actual value, or mean of x, for this particular class, 80.0–84.9, is 14.849, which agrees very closely with the predicted value.

The predicted values obtained by these lines will vary somewhat from the actual values, and this is true especially with the extreme classes in a correlation table where probably only a few individuals are recorded. When a small population is used for determining correlation there may be considerable difference between the predicted and the actual values, and in all cases it must be expected that there will be some variation between the predicted and the actual values. The extent or the amount of the variation expected may be determined by the equation

$$S_x=\sigma_x\sqrt{1-r^2}$$

In this equation S_x is the standard deviation of predicted values, and is termed the standard deviation of an x array.

For the variation in the predicted values of x on y for the data in Table 36, by substituting in the formula we have

$$S_x=1.056\sqrt{1-(.217)^2}=1.031$$

This is the standard deviation of predicted values of x, and means that the predicted values are subject to this much variation. More will be said regarding this point later.

For the variation in predicted values of y on x we have the equation

$$S_y = \sigma_y \sqrt{1 - r^2}$$

In this equation S_y is the standard deviation of predicted values, and is termed the standard deviation of a y array.

From the data in Table 36, using the regression of y on x, it is found that when x equals 13 the predicted value of y is 73.550, while the actual y mean for this particular group is 72.407. For the variation in the predicted values we have

$$S_y = 5.405\sqrt{1 - (.217)^2} = 5.275$$

The predicted value of y, 73.550, comes well within the limits of the variation as expressed by the standard deviation of the predicted value, 5.275.

Using the data in Table 39 and substituting the values of the means, standard deviations, and correlation coefficient in the formula for the regression of x on y, we have

$$X - 18.144 = .938\frac{9.312}{11.472}\,(Y - 23.128)$$
$$X - 18.144 = .938 \times .812\ (Y - 23.128)$$
$$X - 18.144 = .762\ (Y - 23.128)$$
$$X - 18.144 = .762Y - 17.624$$
$$X = .762Y + .520$$

For selected values of y we have the following calculated values of x

For $Y = 11.00$ $\quad X = .762(11) + .520$
$X = 8.382 + .520$
$X = 8.902$

For $Y = 75.00$ $\quad X = .762(75) + .520$
$X = 57.150 + .520$
$X = 57.670$

For $Y=23.128$ $\quad X=.762\,(23.128)+.520$
$X=17.624+.520$
$X=18.144$

For the regression of y on x for the data in Table 39 we have the equation

$$\begin{aligned}
Y-23.128&=\ .938\frac{11.472}{9.312}(X-18.144)\\
Y-23.128&=\ .938\times 1.232\ (X-18.144)\\
Y-23.128&=1.156\ (X-18.144)\\
Y-23.128&=1.156\ X-20.974\\
Y&=1.156\ X+2.154
\end{aligned}$$

For selected values of x the calculated values of y are

For $X=10.00$ $\quad Y=\ 1.156\,(10)+2.154$
$Y=11.560+2.154$
$Y=13.714$

For $X=58.00$ $\quad Y=\ 1.156\,(58)+2.154$
$Y=67.048+2.154$
$Y=69.202$

For $X=18.144$ $\quad Y=\ 1.156\,(18.144)+2.154$
$Y=20.974+2.154$
$Y=23.128$

Here also it is noted that when the mean of x (18.144) is selected the regression line passes through 23.128, the mean of y. When the mean of y (23.128) is selected the calculated value of x is 18.144, the mean of x, showing that the regression lines intersect at the same point where the lines representing the means of x and y intersect. The lines obtained are plotted by connecting the points by means of a straight line, as illustrated in Figure 20, page 128.

As another illustration of the use of regression lines the data from Table 42 are used. Obtaining the means and the proper values for the standard deviations and substituting these values in the equation for the regression of x on y we have

$$\begin{aligned}
X-475.000&=\ .644\frac{140.550}{69.675}(Y-145.225)\\
X-475.000&=\ .644\times 2.017\ (Y-145.225)\\
X-475.000&=1.299\ (Y-145.225)\\
X-475.000&=1.299\ Y-188.647\\
X&=1.299\ Y+286.353
\end{aligned}$$

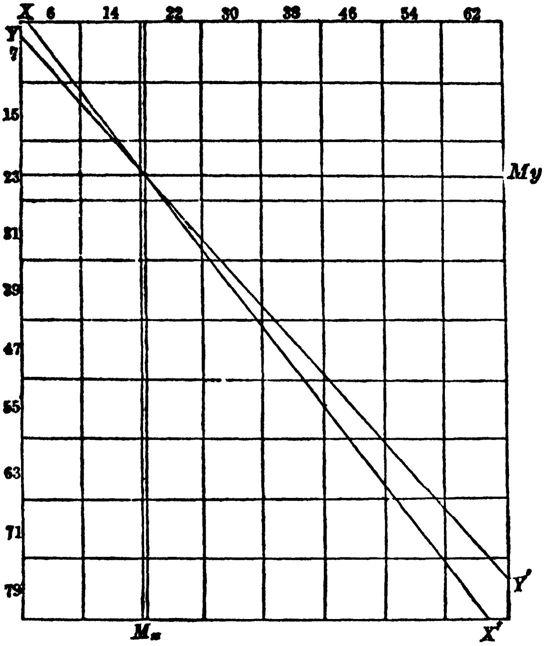


FIG. 20. Regression lines for the data in Table 39. The line XX' shows the regression of x on y and the line YY' shows the regression of y on x.

For selected values of y the calculated values of x are

For $Y=25.0$

$$X=1.299\ (25)+286.353$$
$$X=32.475+286.353$$
$$X=318.828$$

For $Y=300.0$

$$X=1.299\ (300)+286.353$$
$$X=389.700+286.353$$
$$X=676.053$$

For the regression of y on x from the data in Table 42 the equation is

$$Y-145.225=.644\frac{69.675}{140.550}\ (X-475.000)$$
$$Y-145.225=.644\times .496\ (X-475.000)$$
$$Y-145.225=.319\ (X-475.000)$$
$$Y-145.225=.319X-151.525$$
$$Y=.319X-6.300$$

For selected values of x the calculated values of y are

For $X=150.0$
$$\begin{aligned} Y &= .319\ (150.0)-6.300 \\ Y &= 47.850-6.300 \\ Y &= 41.550 \end{aligned}$$

For $X=750.0$
$$\begin{aligned} Y &= .319\ (750.0)-6.300 \\ Y &= 239.250-6.300 \\ Y &= 232.950 \end{aligned}$$

As a special case to show the relation between the predicted and the actual values we may take class 100.0–124.9, and by substituting the mid-point value of this class we find the predicted value of x to be 432.490. The actual mean of x for this class is 431.000. The expected deviation for the predicted value as obtained from the formula

$$S_x=\sigma_x\sqrt{1-r^2}=107.521$$

In this case the predicted and actual values agree very closely.

The meaning of the standard deviation of predicted values may be shown by the scatter diagram, Figure 21, page 130. In this figure each dot represents an individual. For example, in y class 45.0–49.9 there are 2 individuals falling in x class 30.0–39.9. The means, standard deviations, and r have been determined for these data and the regression line RR' for y on x has been obtained. The standard deviation for the predicted values obtained by the formula

$$S_y=\sigma_y\sqrt{1-r^2}=4.733$$

On each side of the regression line RR' the lines SS' are located at a distance of 4.733 from RR'. Other lines may be located at three times the distance ($3S_y$), or 3×4.733, on each side of the line $RR,'$ and it is seen that practically all of the individuals in the table are included. This would be expected from what was stated earlier with reference to the standard deviation, that six times the standard deviation, or three times the standard deviation measured on either side of the mean value (in this case line RR'), includes about 99 per cent of the observations.

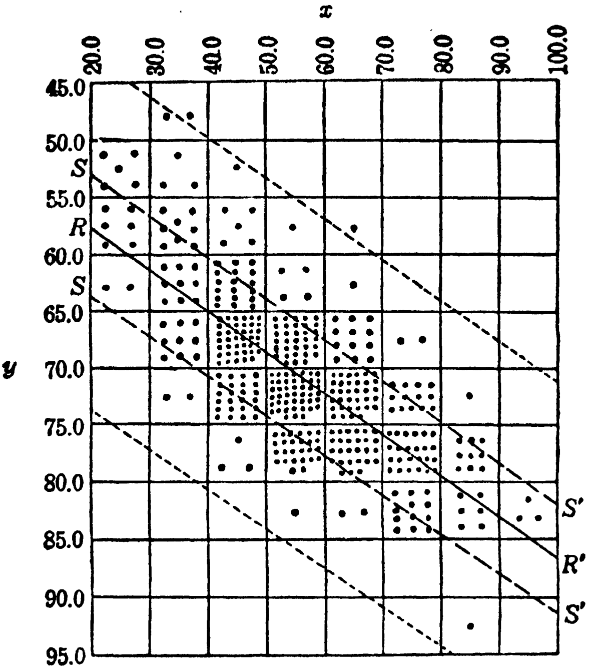


FIG. 21. Scatter diagram showing the relation between two characters of oats. The data are the average number of kernels per culm, x, and the average height of plant in centimeters, y. Each dot represents an individual and the line RR' represents the regression of y on x. On each side of this regression line, lines are drawn at a distance of one times the standard deviation for the predicted values and of three times the standard deviation for the predicted values.

The foregoing illustrates the use of regression lines in predicting expected values. Thus it is possible to predict future events on the basis of past experience, provided sufficient data representing normal conditions have been obtained to justify predictions being made. For example, if over a long period of years the relation of rainfall to the yield of wheat has been determined, then it is possible from the rainfall records of a certain year to predict the wheat yield, keeping in mind of course the standard deviation of the prediction.

Since the correlation coefficient is obtained as a ratio it is therefore independent of the units of measurement, and for this reason it is possible to compare correlation coefficients obtained from different kinds of data. For example, the correlation between weight and height of men may be compared directly with the correlation between the length of right arm and the length of the index finger, although different units of measurement have been used. Further reference will be made in later chapters to the interpretation of correlation coefficients.

CHAPTER VII

SIMPLE CORRELATION—*Continued*

In Chapter VI the measurement of linear correlation was discussed and it was shown how the correlation coefficient, *r*, measures the relationship between two characters. The use of regression lines for predicting expected values for one character on the basis of selected values of a second character was also illustrated. It may often happen that it is desired to determine simple relationships where there may not be sufficient individuals to justify the making of a correlation table but where graphic methods will be satisfactory. Such simple relationships may be shown by means of fitted lines. In Chapter VI a line was fitted merely by observation or inspection (Figure 17). A method for more exact fitting is needed for general use and the theory of least squares furnishes such a method. It is unnecessary here to give the underlying steps leading to the equations for these lines and it is sufficient to state that a line fitted to a series of points by the method of least squares should be such that the sum of the squares of the distances from the several points to the line is a minimum.

Straight Line. The first line that will be discussed is the straight line. To illustrate the fitting of a straight line we may assume the following points: 5, 5, 4, 4, 3, 2, 1, 1. These are shown in the graph in Figure 22.

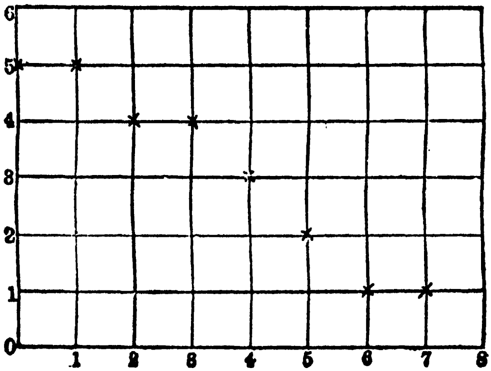


FIG. 22. Showing the points to which a straight line is to be fitted.

The individual points, which we designate the observed y values, are located in the following manner. The first value of y, 5, is located on the first ordinate and is indicated by ×. The next value of y, 5, is located on the second ordinate, and so on for all the points. The method followed here is to take the first ordinate as the starting point, letting the distance between the ordinates be represented by x. The origin for the calculations is at the first ordinate and x for this ordinate is 0. The value of x for the second point is 1, for the next point it is 2, and so on for the different observations. We may if we like take as the starting point one ordinate below the first observation, in which case the value of x for the first point would be 1. As the distance between the ordinates is of equal value, x increases by unity as we proceed from the point of origin, or first ordinate. For fitting straight lines it is desirable to have the data so arranged that the ordinates are separated by equal intervals so that the interval may be taken as unity. When this cannot be done and x does not proceed by equal intervals, the exact interval between each of the ordinates must be used.

The straight line which it is desired to fit to this series of observations may be represented by the equation

$$y=a+bx$$

In this equation y will be the calculated value for any observed value of x; x is taken to represent the distance from the origin for which it is desired to calculate the y value; and a and b are the unknown quantities that it is necessary to obtain in order that a general equation for y may be given. Therefore, when a and b are known, the value of y may be obtained for any selected value of x. The slope of the line will be determined by b. If the sign of b is plus the line slopes upward, and if it is minus the line slopes downward.

Substituting the several observed values in the equation

$$y=a+bx$$

and starting with the first observation, we have

$$
\begin{array}{rl}
 & 5=a+b\ (0) \\
 & 5=a+b\ (1) \\
 & 4=a+b\ (2) \\
 & 4=a+b\ (3) \\
 & 3=a+b\ (4) \\
 & 2=a+b\ (5) \\
 & 1=a+b\ (6) \\
 & 1=a+b\ (7) \\
\text{Total} & 25=8a+28b
\end{array}
$$

Now if each value in these individual equations is multiplied in turn by the coefficient of *b*, or the value of *x* for that equation, we have

$$
\begin{array}{lcl}
5\ (0)=a(0)+b(0)\ (0) & \text{or} & 0 \\
5\ (1)=a(1)+b(1)\ (1) & \text{or} & 5=1a+\ \ 1b \\
4\ (2)=a(2)+b(2)\ (2) & \text{or} & 8=2a+\ \ 4b \\
4\ (3)=a(3)+b(3)\ (3) & \text{or} & 12=3a+\ \ 9b \\
3\ (4)=a(4)+b(4)\ (4) & \text{or} & 12=4a+16b \\
2\ (5)=a(5)+b(5)\ (5) & \text{or} & 10=5a+25b \\
1\ (6)=a(6)+b(6)\ (6) & \text{or} & 6=6a+36b \\
1\ (7)=a(7)+b(7)\ (7) & \text{or} & 7=7a+49b \\
 & \text{Total} & 60=28a+140b
\end{array}
$$

From these two sums we have two equations for *a* and *b*, which when solved will give the required values. It is to be noted that the first equation is obtained by summing the values for *y*, *a*, and *x*. The coefficient of *a* in each individual equation is 1 and summing these we have the coefficient of *a* for the first general equation, and the sum of the *x* values is the coefficient of *b*. The coefficient of *a* equals the number of equations and therefore the first term of the equation may be represented by Σa. This first equation may be transposed and represented by the general equation

$$\Sigma a+\Sigma\ (x)\ b=\Sigma y$$

It will be noted that for the second equation obtained by summation the total of the *x* values equals the coefficient of *a* and the coefficient of *b* is the sum of the squares of *x*, since we multiply the coefficient of *b*, (*x*), in the first equation by *x*. Therefore the second equation may be transposed and expressed in general terms as

$$\Sigma\ (x)\ a+\Sigma\ (x^2)\ b=\Sigma xy$$

These two general equations may therefore be used for fitting straight lines to observed data. It is necessary only to obtain the sum of the x values, the sum of the x^2 values, the sum of y, and the sum of xy. Substituting these in the general equation, we proceed to determine the values for a and b. In actual practice it is not necessary to write down all of the equations as illustrated above and the work may be more easily handled by obtaining the values for the following columns

$$y \quad a \quad x \quad x^2 \quad xy$$

Using the values obtained for the observed data the columns are completed as illustrated in Table 43.

TABLE 43

COLUMNS AND VALUES FOR DETERMINING STRAIGHT LINE FOR ASSUMED DATA

y	a	x	x^2	xy
5	1	0	0	0
5	1	1	1	5
4	1	2	4	8
4	1	3	9	12
3	1	4	16	12
2	1	5	25	10
1	1	6	36	6
1	1	7	49	7
25	8	28	140	60

Substituting in the two general equations the values obtained by summing the columns in Table 43 we have

$$\text{Equation I} \quad 8a + 28b = 25$$
$$\text{Equation II} \quad 28a + 140b = 60$$

It should be noted that to satisfy the conditions for all the points the values for a and b must be calculated from these two general equations rather than from any two of the individual equations. Solving for a and b by the usual method, that is by treating the equations so that the coefficients of a or b are made equal, we have for b

Multiplying Equation I by 7	$56a+196b=\ 175$
Multiplying Equation II by 2	$56a+280b=\ 120$
Subtracting	$-\ 84b=\ 55$
	$b=-.655$

To solve for *a* we may substitute the value for *b* in either of the equations, or solve directly from the two equations as follows:

EQUATION I×5	$40a+140b=\ 125$
EQUATION II	$28a+140b=\ 60$
Subtracting	$12a\ =\ 65$
	$a=5.417$

The equation to this line is therefore

$$y=5.417-.655x$$

The value for *a* determines the origin of the line and since the sign of *b* is minus the line slopes downward, decreasing by the amount of *b* for each added value of *x*. For the second ordinate where $x=1$, the line passes through the point where $y=5.417-(.655\times1)$ or 4.762. Since this line is a straight line two points may be determined and the line drawn through them, and for the last ordinate $y=5.417-(.655\times7)$, or .832. For purposes of checking it is advisable to locate a third point, and if this point falls on the

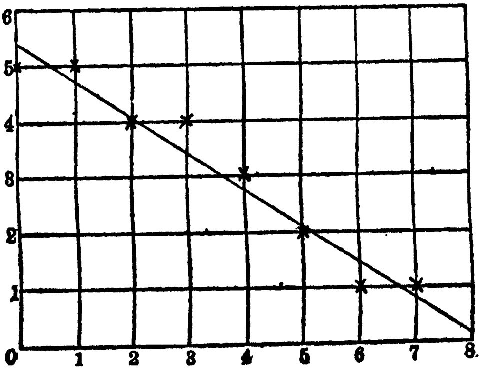


FIG. 23. Showing the fitted line for the points given in Fig. 22.

line it serves as a check on the calculations. The fitted line for this example is shown in Figure 23, page 136.

For a large number of observations the values of x and x^2 may be conveniently obtained from Table I. For larger numbers up to 100 reference may be had to Table XXVIII in *Pearson's Tables for Statisticians and Biometricians.* These summation values may also be obtained directly from the following relations

$$\text{The sum of } N \text{ consecutive numbers} = N\,\frac{(N+1)}{2}$$

$$\text{The sum of the squares of } N \text{ natural numbers} = \frac{N\,(N+1)\,(2N+1)}{6}$$

For the problem above, where it is desired to obtain the sum of the x values from 1 to 7, $N=7$. Substituting in the equation for the sum of N consecutive numbers we have

$$7\left(\frac{8}{2}\right)=28$$

For the sum of the x^2 values, substituting in the equation for the sum of the squares for N natural numbers we have

$$\frac{7\times8\times15}{6}=140$$

These lines are useful in illustrating the trend, and when a line has been fitted to pairs of characters it shows the relationship. This is illustrated in Figure 24, page 138, using the data in the graph in Figure 14 in Chapter VI. This graph shows the mean total yield of cotton according to selected values for the yield of the first picking.

The steps in determining this line, using the columns as suggested, are given in Table 44, page 138, and the explanation following.

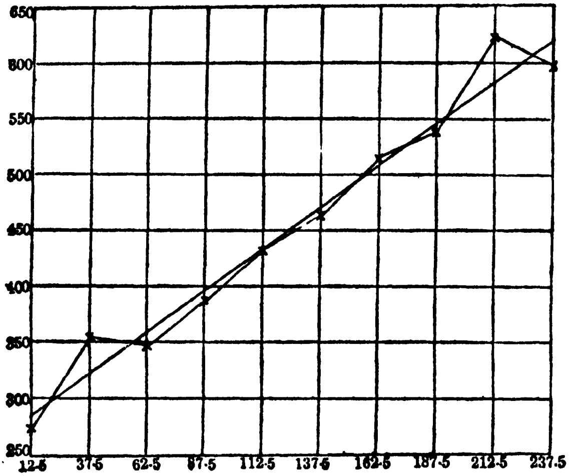


Fig. 24. Straight line fitted to the data illustrated in the graph in Fig. 14. This shows the relation between the yield of cotton at the first picking and the total yield.

TABLE 44

METHOD OF FITTING THE STRAIGHT LINE ILLUSTRATED IN FIGURE 24

MID-POINT VALUE FIRST PICKING	MEAN VALUE TOTAL PICKING y	a	x	x^2	xy
12.5	275.0	1	0	0	0
37.5	356.0	1	1	1	356.0
62.5	344.7	1	2	4	689.4
87.5	385.0	1	3	9	1155.0
112.5	431.0	1	4	16	1724.0
137.5	462.7	1	5	25	2313.5
162.5	514.0	1	6	36	3084.0
187.5	537.5	1	7	49	3762.5
212.5	623.4	1	8	64	4987.2
237.5	595.8	1	9	81	5362.2
TOTAL	4525.1	10	45	285	23433.8

The two equations are

Equation I $10a+45b=4525.1$
Equation II $45a+285b=23433.8$

Solving for a and b

Equation I×4.5 $=45a+202.5b=20362.95$
Equation II $45a+285\ b=23433.8$
Subtracting $-82.5b=-3070.85$
$b=37.222$

Substituting the value for b

Equation I $10a+1674.990=4525.1$
$10a=2850.110$
$a=285.011$

The equation to the line is

$y=285.011+37.222x$
For $x=0$ $y=285.011+37.222\ (0)=285.011$
For $x=9$ $y=285.011+37.222\ (9)=620.009$

It is possible to use straight lines for prediction, as is true of regression lines. The application of straight lines for prediction may be illustrated with the data in Table 45. These are taken from results obtained by the Illinois Agricultural Experiment Station, at Urbana, Illinois, in breeding corn for high oil content, and the observed values in Table 45 are three-year averages.

TABLE 45

APPLICATION OF THE USE OF THE STRAIGHT LINE FOR PREDICTION

YEAR	OBSERVED VALUES	CALCULATED VALUES	PREDICTED VALUES
1898	5.17	5.569	
1901	6.21	6.075	
1904	6.92	6.581	
1907	7.33	7.087	
1910	7.43	7.593	
1913	8.05	8.099	
1916	8.50	8.605	
1919	9.23		9.111
1922	9.96		9.617

Using the results for the first seven periods and obtaining the equation for the straight line, we have

$$y=5.569+.506x$$

The values for the first seven-year periods calculated from this equation may be compared with the observed values and it is seen that there is fairly close agreement. To use the line for prediction we extend the line one more interval to the next three-year period by adding .506 to 8.605, or obtaining

$$y = 5.569 + (.506 \times 7) = 9.111$$

This gives a predicted value of 9.111 which is to be compared with the observed value, 9.23. Using the same argument to predict the average oil content for the next three-year period, the line is extended one more interval and the predicted value is found to be 9.617, which is to be compared with the observed value, 9.96. These predicted values agree fairly well with the observed values.

It must be understood that when using the straight line for prediction in this way we should be certain that the original data used in the calculation of the line represent fairly the facts. That is, if the data are affected by any unusual conditions or are not truly representative we would not be justified in using them to obtain prediction lines.

The use of the straight line may be further illustrated with the data in Table 46, giving the yields of corn at the Illinois Agricultural Experiment Station on the same plots for two consecutive years.

TABLE 46

YIELDS OF CORN FROM THE SAME PLOTS FOR TWO CONSECUTIVE YEARS, USED IN FITTING STRAIGHT LINES ILLUSTRATED IN FIGURES 25 AND 26

DATA AS ORIGINALLY RECORDED			DATA REARRANGED IN ACCORDANCE WITH YIELD FOR 1895		
PLOT No.	YIELD 1895	YIELD 1896	PLOT No.	YIELD 1895	YIELD 1896
101	30.0	87.9	103	25.7	89.8
102	29.1	89.5	104	26.3	94.8
103	25.7	89.8	102	29.1	89.5
104	26.3	94.8	101	30.0	87.9
105	30.3	96.9	105	30.3	96.9
106	31.1	99.5	106	31.1	99.5
107	37.1	94.6	109	34.3	103.9
108	34.6	92.7	108	34.6	92.7
109	34.3	103.9	110	36.3	102.2
110	36.3	102.2	107	37.1	94.6

There are two ways in which the straight line may be used to illustrate the relationship between the yields of the same plots for the different years. One way is to fit a line to the data for each year with the plots in numerical order, as given in Table 46. Such lines have been fitted and are shown in Figure 25. For convenience the scale used for the data for 1895 is indicated on the left of the chart and the scale used for 1896 is indicated on the right.

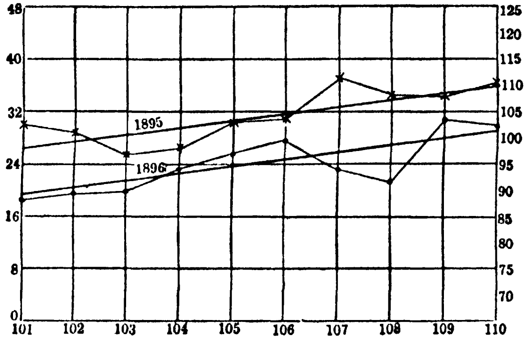


FIG. 25. Straight lines fitted to the data in the second and third columns of Table 46. Since the lines slope in the same general direction there is a relation between the yields of the individual plots for the two years.

The relation between the plots for the different years is indicated by the slope of the two lines, and since they slope in the same general direction it shows that there is a tendency for the plots to respond in the same manner in the two different years.

Another way to apply a straight line to these data is to rearrange the data as given in the second part of Table 46. By this rearrangement, since we are concerned with yields, the yields of the plots for the year 1895 are placed in ascending order. The plot numbers are indicated in the column preceding the yields, but in fitting the line x refers to the difference between the ordinates and not to the difference between plot numbers. By such arrangement the graph for the yields for 1895 in Figure 26, page 142, takes a gradual upward shope, since the data have been arranged in ascending order.

The yields for 1896 are arranged by placing opposite the yield for 1895 the yield of the same plot in 1896. When the data have been rearranged in this way the relation may be shown by fitting only one straight line, which in this case would be fitted to the yields for 1896. Following the usual methods the equation to this line is

$$y=90.774+.979x$$

The fitted line is given in the graph in Figure 26. Since the data for the first year have been arranged in ascending order and since the straight line fitted to the data for the second year also slopes upward, we conclude that there is a relation between the yields of the same plots for the two different years.

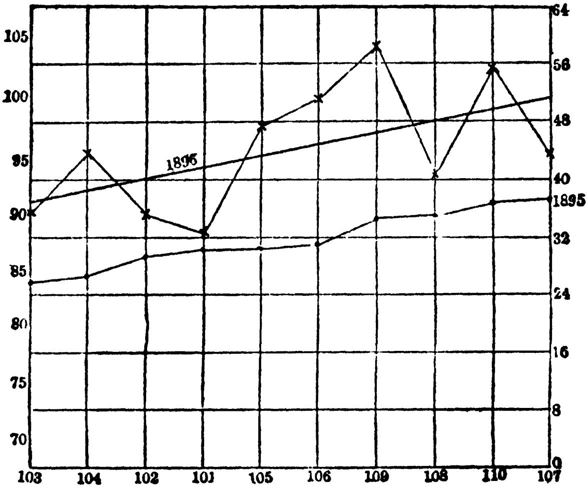


FIG. 26. Straight line fitted to the data in the last column of Table 46. Since the data for 1895 are arranged in order of yield and the straight line slopes in the same general direction, there is a relation between the yields of the individual plots for the two years.

Parabola. Considering further the use of lines to show relationship a straight line has been fitted to the data in Table 47, showing the yields of wheat obtained from different rates of seeding, and the calculated values are given in the third column of the table.

TABLE 47

RESULTS OF FITTING STRAIGHT LINE AND PARABOLAS TO DATA ON YIELDS OF WHEAT OBTAINED FROM DIFFERENT RATES OF SEEDING. DATA FROM RATE AND DATE OF SEEDING TESTS CONDUCTED BY UNITED STATES DEPARTMENT OF AGRICULTURE

RATE OF SEEDING PECKS PER ACRE	SEVEN-YEAR AVERAGE YIELD	VALUES CALCULATED FROM STRAIGHT LINE	VALUES CALCULATED FROM SECOND ORDER PARABOLA	VALUES CALCULATED FROM THIRD ORDER PARABOLA
2	8.9	10.610	9.157	8.758
3	10.3	11.021	10.538	10.669
4	12.3	11.432	11.677	12.006
5	12.9	11.843	12.574	12.865
6	13.1	12.254	13.229	13.342
7	13.5	12.665	13.642	13.533
8	13.8	13.076	13.813	13.534
9	13.6	13.487	13.742	13.411
10	12.7	13.898	13.429	13.350
11	13.5	14.309	12.874	13.357

It is seen that the straight line does not fit the results very well, and this suggests the possibility that a curved line may fit the observed values better than does the straight line. Such a curved line, or parabola, may be fitted to these data by adding to the equation for the straight line a quantity that will give some bend or curvature to the line. The equation for such a line may be represented by

$$y=a+bx+cx^2$$

This is referred to as a second order parabola, considering the straight line as the first order parabola. In this equation there are three unknowns, a, b, and c. When these values are determined y may be calculated for any selected value of x. It is not necessary to give the steps leading to the different equations, but since we have three unknowns we must have three equations for determining the values of a, b, and c. Following the plan for the straight line these three equations may be expressed in general terms as

EQUATION I $\Sigma a+\Sigma(x)b+\Sigma(x^2)c=\Sigma y$
EQUATION II $\Sigma(x)a+\Sigma(x^2)b+\Sigma(x^3)c=\Sigma xy$
EQUATION III $\Sigma(x^2)a+\Sigma(x^3)b+\Sigma(x^4)c=\Sigma x^2y$

Following the system suggested for the straight line, columns may be arranged as follows:

$y \quad a \quad x \quad x^2 \quad x^3 \quad x^4 \quad xy \quad x^2y$

For the data in Table 47 the columns are

y	a	x	x^2	x^3	x^4	xy	x^2y
8.9	1	0	0	0	0	0	0
10.3	1	1	1	1	1	10.3	10.3
12.3	1	2	4	8	16	24.6	49.2
12.9	1	3	9	27	81	38.7	116.1
13.1	1	4	16	64	256	52.4	209.6
13.5	1	5	25	125	625	67.5	337.5
13.8	1	6	36	216	1296	82.8	496.8
13.6	1	7	49	343	2401	95.2	666.4
12.7	1	8	64	512	4096	101.6	812.8
13.5	1	9	81	729	6561	121.5	1093.5
124.6	10	45	285	2025	15333	594.6	3792.2

Substituting these summation values in the three general equations we have

EQUATION I $10a+45b+285c=124.6$
EQUATION II $45a+285b+2025c=594.6$
EQUATION III $285a+2025b+15333c=3792.2$

Solving from these equations for the unknown quantities a, b, and c we have

$$
\begin{aligned}
a &= 9.157\\
b &= 1.502\\
c &= -.121
\end{aligned}
$$

Substituting these values in the equation for the second order parabola the equation becomes

$$y=9.157+1.502x-.121x^2$$

The calculated values from the second order parabola for the various rates of seeding are given in column 4 of Table 47. The values obtained from this parabola fit the observed values better than do those obtained from the straight line.

It is noted that where the rate of seeding is 10 pecks there is a slight drop in the observed yield, 12.7, as compared with the yield obtained for the rate of 11 pecks of seed. This drop is not to be expected from the nature of the case, nevertheless for purposes of illustration these data are used to show how such points may be better fitted by means of a parabola of higher order. This is done by adding another term to the equation for the second order parabola, giving the equation for the third order parabola

$$y=a+bx+cx^2+dx^3$$

In this equation there are four unknown values, a, b, c, and d, whose values are to be determined. It may be noted here that while the different equations are really made up by adding an additional term to the preceding equation, as for example with the equation to the straight line, $y=a+bx$, a third term, cx^2, is added to give the equation for the second order parabola, the numerical values of a and b will necessarily be different in the two equations.

Since there are four unknowns in the equation for the third order parabola it is necessary to have four general equations, which may be stated

EQUATION I $\Sigma a+\Sigma(x)b+\Sigma(x^2)c+\Sigma(x^3)d=\Sigma y$
EQUATION II $\Sigma(x)a+\Sigma(x^2)b+\Sigma(x^3)c+\Sigma(x^4)d=\Sigma xy$
EQUATION III $\Sigma(x^2)a+\Sigma(x^3)b+\Sigma(x^4)c+\Sigma(x^5)d=\Sigma x^2y$
EQUATION IV $\Sigma(x^3)a+\Sigma(x^4)b+\Sigma(x^5)c+\Sigma(x^6)d=\Sigma x^3y$

For solving these equations it would be necessary to have the following columns

$$y \quad a \quad x \quad x^2 \quad x^3 \quad x^4 \quad x^5 \quad x^6 \quad xy \quad x^2y \quad x^3y$$

By substituting the several values as before the summations obtained for these columns are

$y=124.6$
$x=45$
$x^2=285$
$x^3=2025$
$x^4=15333$
$x^5=120825$
$x^6=978405$
$xy=594.6$
$x^2y=3792.2$
$x^3y=26972.4$

By substituting these values in the four general equations we have

EQUATION I $10a+45b+285c+2025d=124.6$
EQUATION II $45a+285b+2025c+15333d=594.6$
EQUATION III $285a+2025b+15333c+120825d=3792.2$
EQUATION IV $2025a+15333b+120825c+978405d=26972.4$

The values for *a*, *b*, *c*, and *d* determined from these equations are

$a=8.758$
$b=2.230$
$c=-.335$
$d=.016$

and the equation is

$$y=8.758+2.230x-.335x^2+.016x^3$$

By using these values for the several values of x we have the results as given in column 5 of Table 47.

By the addition of another term to the general equation the calculated value for the 10-peck rate shows a slight drop as compared with the value for the 11-peck rate. This is in accordance with the original data, and while the original value is no doubt purely accidental the data are used here merely to show how parabolas of higher order may be useful in fitting values to data which by nature would show such tendencies. It is possible to fit parabolas of higher order by the continual addition of another term, as, for example, the equation for the next higher order parabola would be

$$y=a+bx+cx^2+dx^3+ex^4$$

The calculation of parabolas of the higher orders becomes rather laborious, and it is doubtful whether it is often worth while to carry the calculations beyond possibly the third or fourth order parabola.

Correlation Ratio. From the results of the application of the fitted lines it is clear that the relationship may at times be shown better by means of a curved line rather than a straight line. This is true with many problems in correlation when we have what is termed non-linear correlation.

The correlation coefficient and the regression lines, as discussed in Chapter VI, measure the correlation and predict expected results on the assumption that the means of the rows and the means of the columns of a correlation table fall on a straight line. Therefore it is assumed that when using this measure for determining the correlation or the regression line based on the value of r we are dealing with cases of linear correlation or of linear regression. The equations which were used to measure the standard deviations of arrays

$$S_x=\sigma_x\sqrt{1-r^2}$$

$$S_y=\sigma_y\sqrt{1-r^2}$$

really measure the scatter for each array about the regression line, which for linear correlation is a straight line. Let us now square and transpose each of these equations and we have

$$r^2=\frac{\sigma^2_x-S^2_x}{\sigma^2_x}$$

and

$$r^2=\frac{\sigma^2_y-S^2_y}{\sigma^2_y}$$

and from these

$$r_{xy}=\sqrt{1-\frac{S^2_x}{\sigma^2_x}}$$

and

$$r_{yx}=\sqrt{1-\frac{S^2_y}{\sigma^2_y}}$$

Thus we have the measure of the correlation coefficient, r, in terms of the ratio of the scatter of x or y to the standard deviation for the entire distribution of x or the entire distribution of y. The first equation gives the correlation of x on y and the second of y on x. These equations may be used for a general measure for correlation where S_x or S_y represents the scatter about the best fitting line, and therefore they may be used for linear or non-linear correlation. For linear correlation the numerical values would be the same.

Since in many instances the relationship may be better represented by curved lines rather than straight lines, the values obtained in measuring the scatter from a straight line will be larger than the values obtained from curved lines that will better fit the means of the rows or of the columns. Since the correlation coefficient, r, is calculated on the assumption that the means of the rows fall in a straight line it means that the value of r will be lower numerically than would be the case if the scatter of deviations were measured from the curved line that would best fit the means of the rows or the means of the columns. In order, then, to have a general measure for correlation which will be independent of the type of line, Pearson has suggested the correlation ratio as a measure of such correlation, and this is represented by the Greek letter Eta, η.

Using the above equations for r we may now write

$$\eta_{xy}=\sqrt{1-\frac{S^2{}_x}{\sigma^2{}_x}}$$

and

$$\eta_{yx}=\sqrt{1-\frac{S^2{}_y}{\sigma^2{}_y}}$$

From the first equation we have the correlation ratio of x on y and from the second equation the correlation ratio of y on x, and for linear correlation $r=\eta$. It is clear that there are two values of η for each correlation table, while there is only one value for r.

For purposes of calculation the equations for η_{xy} and η_{yx} may be put in the following form

$$\eta_{xy}=\frac{\sqrt{\frac{\Sigma n_y\,(m_x-M_x)^2}{N}}}{\sigma_x}$$

$$\eta_{yx}=\frac{\sqrt{\frac{\Sigma n_x(m_y-M_y)^2}{N}}}{\sigma_y}$$

In these equations n_y represents the total frequency of any row and n_x the total frequency of any column; m_x is the mean of any row and m_y is the mean of any column; M_x is the mean of the entire x population and M_y the mean of the entire y population; N equals the total number in the population; and σ_x is the standard deviation for the entire x distribution and σ_y is the standard deviation for the entire y distribution.

We may now write

$$\sigma_{m_x}=\sqrt{\frac{\Sigma n_y(m_x-M_x)^2}{N}}$$

$$\sigma_{m_y}=\sqrt{\frac{\Sigma n_x(m_y-M_y)^2}{N}}$$

We may then write the formulas for the correlation ratios as

$$\eta_{xy}=\frac{\sigma_{m_x}}{\sigma_x}$$

and

$$\eta_{yx}=\frac{\sigma_{m_y}}{\sigma_y}$$

The steps in calculating these two values for the correlation ratio are given in Table 48, page 150. We proceed by taking each row or column in turn and determining the difference between the mean of the row or of the column and the mean of the entire population for the same character. These differences are then squared and multiplied by their respective frequencies. This is done for all of the rows or for all of the columns. When the difference between the mean of each row or column and the mean of the whole population is squared and multiplied by its frequency, we have the weighted-squared deviations of the distribution of the rows or of the columns. The sum of these values is divided by the number in the population to obtain the mean-weighted-squared deviation. Extracting the square root of this

TABLE 48

METHOD OF CALCULATING THE CORRELATION RATIO. DATA ARE THE AVERAGE HEIGHT OF PLANT, x, AND THE AVERAGE YIELD OF CULM PER PLANT, y, IN OATS

	x											f	m_x	$m_x - M_x$ or D	D^2	fD^2
	25.0 29.9	30.0 34.9	35.0 39.9	40.0 44.9	45.0 49.9	50.0 54.9	55.0 59.9	60.0 64.9	65.0 69.9	70.0 74.9	75.0 79.9					
0.0–0.9	1	10	25	19	19	4						78	41.154	−13.829	191.241	14916.798
1.0–1.9	1				20	40	28	6	1			96	53.437	− 1.546	2.390	229.440
2.0–2.9						3	22	29	16	2		72	61.944	6.961	48.456	3488.832
y 3.0–3.9								11	22	4		37	66.554	11.571	133.888	4953.856
4.0–4.9									5	5	2	12	71.250	16.267	264.615	3175.380
5.0–5.9										2	2	4	75.000	20.017	400.680	1602.720
6.0–6.9											1	1	77.500	22.517	507.015	507.015
																28874.041
f	2	10	25	19	39	47	50	46	44	13	5	300				

f	m_y	$m_y - M_y$ or D	D^2	fD^2
2	1.000	− .917	.841	1.682
10	.500	−1.417	2.008	20.080
25	.500	−1.417	2.008	50.200
19	.500	−1.417	2.008	38.152
39	1.013	− .904	.817	31.863
47	1.479	− .438	.192	9.024
50	1.940	.023	.001	.050
46	2.609	.692	.479	22.034
44	3.205	1.288	1.659	72.996
13	4.038	2.121	4.499	58.487
5	5.300	3.383	11.445	57.225
300				361.793

$M_x = 54.983$
$\sigma_x = 10.950$

$$\sigma_{m_x} = \sqrt{\frac{28874.041}{300}} = 9.811$$

$$\eta_{xy} = \frac{9.811}{10.950} = .896$$

$r = .865$

$M_y = 1.917$
$\sigma_y = 1.215$

$$\sigma_{m_y} = \sqrt{\frac{361.793}{300}} = 1.098$$

$$\eta_{yx} = \frac{1.098}{1.215} = .904$$

quantity gives the standard deviation of the means of the rows or the means of the columns, depending on which values have been used. The ratio of this standard deviation of the means of the rows or columns to the standard deviation of the total frequency of x or y gives the correlation ratio.

We may illustrate these various steps by taking the first y array for class 0.0–0.9 in Table 48. The mean of this array for x is obtained, using the mid-points of the x classes and making a frequency distribution for the y array in the following manner:

V	f	fV
27.5	1	27.5
32.5	10	325.0
37.5	25	937.5
42.5	19	807.5
47.5	19	902.5
52.5	4	210.0
	78	3210.0

3210.0/78 = 41.154

The short method for obtaining the mean from a frequency distribution could also be used. The deviation between the mean of this array, 41.154, and the mean of the entire x population, 54.983, is obtained, squared, and multiplied by the frequency of the array, 78. Each of the arrays is handled in the same manner. Completing the calculations we obtain $\sigma_{m_x} = 9.811$. Substituting this value in the equation

$$\eta_{xy} = \frac{\sigma_{m_x}}{\sigma_x}$$

we have

$$\eta_{xy} = \frac{9.811}{10.950} = .896$$

In a similar manner the correlation ratio for y on x is obtained from

$$\eta_{yx} = \frac{\sigma_{m_y}}{\sigma_y}$$

and

$$\eta_{yx} = \frac{1.098}{1.215} = .904$$

It is seen that both values for the correlation ratio are higher numerically than the correlation coefficient, .865.

If there were no scatter at all then the standard deviation in the numerator and the standard deviation in the denominator would be equal, and therefore the correlation ratio would be 1. Since the correlation ratio is the result obtained by dividing two standard deviations, the correlation ratio does not show whether the relationship is positive or negative. In other words, when the correlation ratio is obtained it does not show by its sign, as does the correlation coefficient, whether there is a positive or a negative relationship between the two characters.

In the discussion in the previous chapter relative to making correlation tables it was pointed out that such tables are desirable for presenting the data graphically. Since it may be possible that we are dealing with a case of non-linear relationship, if we determine the correlation coefficient, r, without grouping the full degree of the association may not be represented by the numerical value of r. If we have a number of pairs of characters it is difficult to tell by mere inspection whether we are dealing with a case of linear or non-linear relationship, while from a correlation table we obtain a better idea of the kind of relationship with which we are dealing.

Another illustration showing the difference between the correlation coefficient and the correlation ratio is given in Table 49, page 153. Following the same methods for determining the correlation ratios as in the preceding table we find

$$\eta_{xy} = .647 \text{ and } \eta_{yx} = .714$$

From both of these illustrations it is seen that the correlation values for a correlation table are not numerically the same. This may be expected, but in both cases the correlation ratio values are higher than the correlation coefficient. In some instances the correlation ratio may be about the same value as the correlation coefficient, as in Table 49 where η_{xy} is .647 and the correlation coefficient is .635. This may mean that one of the characters in the correlation does not deviate from linearity while the other character does. In all cases it is better to determine both values for the correlation ratio.

TABLE 49

RELATION BETWEEN YIELD IN GRAMS, x, AND NUMBER OF BRANCHES PER PLANT, y, IN BUCKWHEAT

	x																
y	2	6	10	14	18	22	26	30	34	38	42	46	50	54	58	62	f
0	4																4
1	1																1
2	6																6
3	8	7	3	1													19
4	2	8	12	7	5	2	4	1									41
5	2	9	15	17	23	13	8	5	1	1	1	1					96
6	1	3	11	19	21	23	23	6	9	3	4	6		1	2		132
7	1		2	11	15	21	22	21	12	7	8	4	2	3			129
8			2	1	2	4	7	11	11	8	5	1	7	3	2		64
9						1		2	2	1	3	1			1	1	12
10						1			1		1						3
11										1							1
	25	27	45	56	66	65	64	46	36	21	22	13	9	7	5	1	508

$r = .635$ $\eta_{xy} = .647$ $\eta_{yx} = .714$

The correlation ratio may be compared with the correlation coefficient to determine whether the difference is large enough to be significant. This difference depends on the difference between the squares of the correlation ratio and the correlation coefficient, or in other words the difference between these two, η^2-r^2, is the test of linearity. If the relationship is a linear one then these two values, η^2 and r^2, will be equal, and the more the relationship diverges from linearity the greater will be the difference between the squares of η and r. A method which has been frequently used to determine the significance of this difference, but which is not exact since no account is taken of the number of arrays, is given by the following equation:

$$\sigma_{(\eta^2-r^2)}=2\sqrt{\frac{\eta^2-r^2}{N}}$$

Using in this equation the values for the regression of y on x from Table 48, $\eta_{yx}=.904$ and $r=.865$, we have

$$\sigma_{(\eta^2-r^2)}=2\sqrt{\frac{(.904)^2-(.865)^2}{300}}=.030$$

The difference between η^2 and r^2 is .069. Usually unless this difference is three times the standard deviation of the difference it is understood that no real difference exists between the two values. In other words, for this particular case, while the correlation ratio is a little higher than the correlation coefficient, the difference is not significant.

Using the value for the correlation ratio for y on x from Table 49, where η_{yx} is .714 and the correlation coefficient is .635, the difference between η^2 and r^2 is .107 and the standard deviation of the difference is .029. In this case the difference is more than three times the standard deviation of the difference, meaning that the difference is significant and that the correlation ratio is the better measure for the correlation between these two characters. More will be said regarding the interpretation of this difference in Chapter XIII. It should be clear that the correlation ratio is a more general measure of correlation, and even in cases where the deviation from linearity is slight it gives a more exact measure of correlation.

Pearson has suggested a correction for the correlation ratio as follows:

$$\text{Corrected } \eta^2 = \frac{\eta^2 - \frac{(K-3)}{N}}{1 - \frac{(K-3)}{N}}$$

in which K represents the number of arrays. Thus, from Table 48, for η_{xy}, we have

$$\eta_{xy} = .896$$

$$\begin{aligned}\text{Corrected } \eta^2 &= \frac{(.896)^2 - \frac{7-3}{300}}{1 - \frac{7-3}{300}}\\ &= \frac{.802816 - .013333}{1 - .013333}\\ &= .800151\\ \text{Corrected } \eta &= \sqrt{.800151} = .895\end{aligned}$$

The correction is small, but for a small population or for many arrays the correction would be larger.

Curved Regression Lines. In cases of linear relationship regression lines are calculated in order that they may be used for predicting expected values in one character for selected values in the other. In the case of non-linear correlation we also have regression lines, but these are curved regression lines. Fisher has given a means of fitting curved regression lines, and Tippett has followed Fisher in showing the application. We will now illustrate the fitting of curved regression lines according to this method. Following the notation of Tippett the general equation for such regression lines may be written

$$Y = a + bt + ct^2 + dt^3$$

Let y = the mean of an array for a correlation table, or for ungrouped data it may be the value of any item.

Let t = the deviation of any array from the middle array. If the number of arrays are even then t is still the deviation from the middle, which then becomes the average of the two central arrays. For ungrouped data the middle items are the ones from which t is measured.

The equation may be transformed to

$$Y = A + BT_1 + CT_2 + DT_3$$

in which

$$T_1 = (t - M_t) = t, \text{ since } M_t, \text{ or the mean of } t, = 0$$

$$T_2 = t^2 - \frac{N^2-1}{12}$$

$$T_3 = t^3 - \frac{3N^2-7}{20}t$$

The values of A, B, C, and D may be found from the following relations

$$A = \frac{\Sigma(y)}{N}$$

$$B = \frac{12}{N(N^2-1)}\Sigma(yT_1)$$

$$C = \frac{180}{N(N^2-1)(N^2-4)}\Sigma(yT_2)$$

$$D = \frac{2800}{N(N^2-1)(N^2-4)(N^2-9)}\Sigma(yT_3)$$

For purposes of calculation, B, C, and D may be more conveniently obtained by the following equations, where N is the number of arrays:

$$B = \frac{12}{N(N^2-1)}\Sigma(yt)$$

$$C = \frac{180}{N(N^2-1)(N^2-4)}\left[\Sigma(yt^2) - \left(\frac{N^2-1}{12}\right)\Sigma(y)\right]$$

$$D = \frac{2800}{N(N^2-1)(N^2-4)(N^2-9)}\left[\Sigma(yt^3) - \left(\frac{3N^2-7}{20}\right)\Sigma(yt)\right]$$

If a curve of the next higher order is needed then

$$T_4 = t^4 - \frac{3N^2-13}{14}t^2 + \frac{3(N^2-1)(N^2-9)}{560}$$

and

$$E = \frac{44100}{N(N^2-1)(N^2-4)(N^2-9)(N^2-16)}\Sigma(yT_4)$$

The data for the y means for the x arrays in Table 48 may now be taken to illustrate the fitting of a line by use of these equations. The origin for the calculations is taken at the middle class and the deviations of the several classes are marked off by intervals of 1.

The deviations below the middle class are preceded by a minus sign. It may be stated that if the number of classes were even the origin would be taken at the center and the values of t would be .5, 1.5, 2.5, and so on. The following columns are arranged and the necessary sums of these columns obtained.

$$y \quad t \quad t^2 \quad t^3 \quad yt \quad yt^2 \quad yt^3$$

For the y means for the x arrays in Table 48 the values in these columns would be as follows:

y	t	t^2	t^3	yt	yt^2	yt^3
1.00	-5	25	-125	-5.00	25.00	-125.00
.50	-4	16	- 64	-2.00	8.00	- 32.00
.50	-3	9	- 27	-1.50	4.50	- 13.50
.50	-2	4	- 8	-1.00	2.00	- 4.00
1.01	-1	1	- 1	-1.01	1.01	- 1.01
1.48	0	0	0	0	0	0
1.94	1	1	1	1.94	1.94	1.94
2.61	2	4	8	5.22	10.44	20.88
3.20	3	9	27	9.60	28.80	86.40
4.04	4	16	64	16.16	64.64	258.56
5.30	5	25	125	26.50	132.50	662.50
TOTAL 22.08				48.91	278.83	854.77

Substituting these summation values in the equations for A, B, C, and D we obtain

$$\begin{aligned} A &= 2.007 \\ B &= .445 \\ C &= .068 \\ D &= -\ .003 \end{aligned}$$

These values will be used in the general equation

$$Y = A + BT_1 + CT_2 + DT_3$$

and in accordance with the different values of t the calculated points will be determined for each array. For example, where $t = -5$, we have

$$T_1 = -5$$

$$T_2 = 25 - \frac{121-1}{12} = 15$$

$$T_3 = -125 - \frac{363-7}{20} \times -5 = -36.0$$

and in the general equation, for this particular array

$$\begin{aligned} Y&=2.007+(.445\times -5)+(.068\times 15)-(.003\times -36.0)\\ &=2.007-2.225+1.020+.108=.910 \end{aligned}$$

Where

$t=5$, we have

$$T_1=5$$

$$T_2=25-\frac{121-1}{12}=15$$

$$T_3=125-\frac{363-7}{20}\times 5=36.0$$

In the general equation, for this array

$$\begin{aligned} Y&=2.007+(.445\times 5)+(.068\times 15)-(.003\times 36.0)\\ &=2.007+2.225+1.020-.108=5.144 \end{aligned}$$

The calculated values for the y means for the x arrays from Table 48 are shown in the last column in Table 50. In making the calculations from these various equations it is very important to observe the proper signs.

TABLE 50

RESULTS OF FITTING DIFFERENT LINES TO DATA IN TABLE 48, ON AVERAGE YIELD OF CULM PER PLANT FOR THE x ARRAYS

AVERAGE YIELD OF CULM PER PLANT FOR x ARRAYS	VALUES CALCULATED FROM STRAIGHT LINE	VALUES CALCULATED FROM SECOND ORDER PARABOLA	VALUES CALCULATED FROM THIRD ORDER PARABOLA	VALUES CALCULATED BY METHOD FOR CURVED REGRESSION LINE
1.00	-.216	.798	.892	.910
.50	.229	.635	.616	.613
.50	.674	.608	.536	.525
.50	1.119	.717	.634	.626
1.01	1.564	.962	.892	.900
1.48	2.009	1.343	1.292	1.327
1.94	2.454	1.860	1.816	1.890
2.61	2.899	2.513	2.446	2.572
3.20	3.344	3.302	3.164	3.353
4.04	3.789	4.227	3.952	4.217
5.30	4.234	5.288	4.792	5.144

Applying this method the curved regression line may be calculated for the x means for the y arrays from Table 48 following the equations just given, and the values obtained for A, B, C, and D are

$$A = 63.829$$
$$B = 5.764$$
$$C = -.859$$
$$D = .150$$

These are substituted in the general equation and the calculated values for y are obtained. These calculated values for y are shown in Table 51, with the observed values.

TABLE 51

RESULTS OF FITTING CURVED REGRESSION LINE TO DATA IN TABLE 48, ON AVERAGE HEIGHT OF CULM PER PLANT FOR THE y ARRAYS

AVERAGE HEIGHT OF CULM PER PLANT FOR y ARRAYS	VALUES CALCULATED FROM CURVED REGRESSION LINE
41.2	41.342
53.4	53.201
61.9	61.542
66.6	67.265
71.2	71.270
75.0	74.457
77.5	77.726

The regression lines fitted by this method for the data in Table 48 are shown in Figure 27, page 160.

While this method gives better results when the original values are of equal weight, very good approximations are obtained when it is applied to unweighted values as in a correlation table. The method may also be followed where the correlation has been worked without grouping, or applied to a series of observations in place of the parabola. In the latter application the origin is taken at the center and if there are several observations, as for example the results over a period of years, the difference between any one year and the middle year will be designated by t, and the calculations carried out in the usual way.

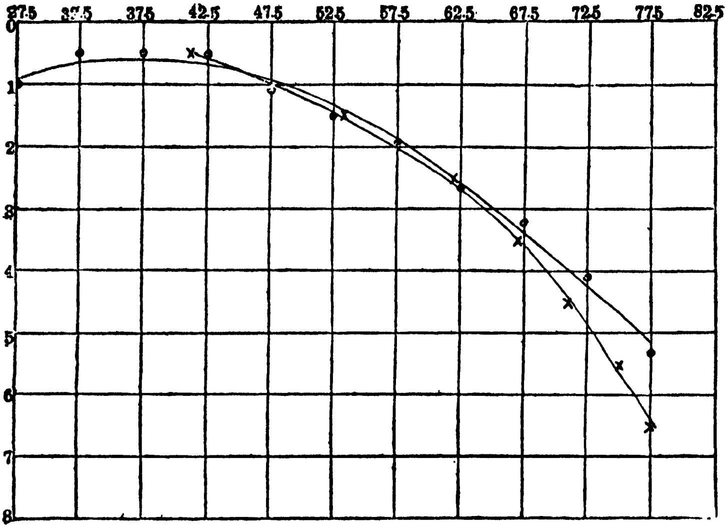


FIG. 27. Curved regression lines fitted to the data in Table 48. The dots represent the average yield of culm and the crosses represent the average height of plant.

By inspection of Table 50 it is seen that the values calculated by this method for the curved regression line fit the observed values much better than do those from the other lines calculated. It is evident that the regression is not linear and the straight line does not fit the observations. The values obtained from either the second or third order parabolas fit better than do those from the straight line, but not so well as do those obtained from the method for the curved regression line.

These curved regression lines are used for prediction in the same manner as the normal regression lines. When applied to a correlation table the scatter or standard deviation for any row or column would be determined by the usual formula for measuring the scatter.

This method of fitting curved regression lines has one distinct advantage in that for determining a higher order parabola the values for the additional term may be added to the equation without changing the values of the preceding terms. On the other hand, when calculating parabolas from the formula $y=a+bx+cx^2 \dots$ the values of the preceding terms change as a new term is added.

Correlation from Ranks. The methods just described determine the correlation by considering the position and value of each item in the series. It is often convenient to determine the correlation by considering the position only, or the rank, of each item. This may be done with less labor than is required for the usual methods, and is especially useful for making a rapid determination of the amount of correlation. Spearman has developed a method for obtaining the correlation from the rank or position. By this method, to determine the correlation between two series, x and y, a rank is assigned to each value of x and another rank to each value of y. These ranks are determined by noting the values of each item, and beginning with the highest or lowest value the ranks are assigned in order. The correlation may then be determined from these ranks.

As an illustration of this method we may use the data in Table 52. These are the yields of the same varieties of wheat under two different systems of planting, three-row plots and single-row plots.

TABLE 52

METHOD OF CALCULATING CORRELATION FROM RANKS

THREE-ROW PLOTS (1)	SINGLE-ROW PLOTS (2)	RANK OF (1) x	RANK OF (2) y	$x-y$ or D	D^2
36.7	34.1	6.5	7	− .5	.25
42.5	40.6	2	3	−1.0	1.00
44.9	41.0	1	2	−1.0	1.00
41.0	42.4	3	1	2.0	4.00
32.6	31.7	9	8	1.0	1.00
40.0	36.5	5	5	0	0
40.5	36.4	4	6	−2.0	4.00
36.7	37.6	6.5	4	2.5	6.25
34.0	31.6	8	9	−1.0	1.00
31.3	28.1	11	12	−1.0	1.00
32.2	30.7	10	10	0	0
31.2	29.9	12	11	1.0	1.00
29.6	23.7	13	13	0	0
					$\Sigma D^2 = 20.50$

$$r_r = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

$$r_r = 1 - \frac{6\times 20.50}{13(169-1)} = 1 - \frac{123.00}{2184} = .944$$

The varieties are ranked in accordance with their yield and, as stated, we may begin with either the highest or lowest yield. In this case the ranking is made beginning with the highest yield. Thus, plot number 3 in the three-row plots has the highest yield and it is given rank number 1. Plot number 2 is given rank number 2. When two or more observations have the same numerical value the ranks for those values are divided between them. For example, in the three-row plots in Table 52 there are two plots having a yield of 36.7 bushels. As this value comes immediately after the value for rank number 5 it would normally be given rank number 6, but since there are two values exactly the same they are given the average of the ranks 6 and 7, and each is assigned the rank of 6.5. This method would be followed if more than two of the same values were obtained. For example, if there had been three values of 36.7, then the average of 6, 7, and 8, or 7, would be the rank assigned to each of the three.

The yields for the second system of planting are arranged by rank in a similar manner. The differences between the ranks of the two systems of planting are then obtained, squared, and summed. The correlation coefficient obtained in this way will be designated r_r, and the formula is

$$r_r = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

Here N refers to the number of items. Substituting in this formula the values obtained in Table 52 the correlation is .944.

This is a convenient way of obtaining correlation if the number of items is not too large, but this method should not be substituted for the usual methods of determining correlation by means of the correlation coefficient and the correlation ratio. It will give a measure for correlation that approximates closely the value obtained for the correlation coefficient.

Another illustration of the use of the rank method for determining correlation is given in Table 53 on page 163. The data in this table are the results obtained from a yield comparison of a number of wheat varieties grown in three-row plots.

TABLE 53

ANOTHER ILLUSTRATION OF THE METHOD OF CALCULATING CORRELATION FROM RANKS

AVE. ALL ROWS	AVE. MIDDLE ROWS	RANK OF ALL ROWS x	RANK OF MIDDLE ROWS y	$x-y$ OR D	D^2
47.0	53.6	5	2	3	9
27.6	31.3	19	17	2	4
41.4	47.1	10	8	2	4
36.9	45.0	13	9	4	16
11.5	12.4	27	26	1	1
49.8	53.0	4	3	1	1
29.0	32.8	16	16	0	0
22.4	27.3	23	19	4	16
45.1	48.8	7	6	1	1
12.1	12.1	26	27	-1	1
52.9	54.7	1	1	0	0
25.5	25.2	21	23	-2	4
17.0	15.9	24	25	-1	1
36.3	38.3	14.5	13	1.5	2.25
36.3	37.6	14.5	14	.5	.25
46.0	47.5	6	7	-1	1
51.9	51.9	2	5	-3	9
15.6	16.1	25	24	1	1
28.3	26.5	18	20	-2	4
23.1	25.3	22	22	0	0
28.4	28.3	17	18	-1	1
26.8	25.8	20	21	-1	1
38.6	37.5	11	15	-4	16
38.4	40.3	12	12	0	0
8.2	7.2	28	28	0	0
4.8	4.3	29	29	0	0
51.8	52.0	3	4	-1	1
42.3	43.5	8	10	-2	4
42.2	43.0	9	11	-2	4
					$\Sigma D^2 = 102.50$

The formula is $r_r = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$

$$r_r = 1 - \frac{6\times 102.50}{29\ (841-1)} = 1 - \frac{615.00}{24360} = .975$$

The data in the first column of the table are the averages of all three rows and the data in the second column are the averages of the middle rows for the same varieties. Ranks are assigned in the manner explained and the correlation is found to be .975.

As stated above, the value obtained for correlation by the rank method will approximate the value of the correlation coefficient. It is possible to make a correction for the value obtained from the rank method by the following relation:

$$r=2\,sin\left(\frac{\pi r_r}{6}\right)$$

For the correlation just determined $r_r=.975$, and

$$r=2\,sin\left(\frac{\pi .975}{6}\right)$$

in which π is taken as 180° (180° $=\pi$ radians). Thus we have

$$\frac{180\times .975}{6}=29.25^{\circ}$$

Now, from tables of trigonometric functions the logarithm of *sin* 29.25° is 9.68897–10. The number corresponding to this logarithm is found from a table of common logarithms to be .4886. Multiplying .4886 by 2 we have .9772, which is the value of r for this series of observations computed from r_r. Table III in the Appendix gives corresponding values for r computed from r_r.

Spearman has also given another method for determining r, called the foot-rule method. This method gives only a very rough estimate of correlation and is not directly comparable with other measures of correlation. It may be useful as a quick means of obtaining some idea of the relationship. It is based on the gain of the rank of one character over the rank of the second character. Using the data in Table 53 the ranks are assigned, and where the rank in the second method of planting is higher than in the first method of planting the difference is recorded. Where the rank of the second method is equal to or below that of the first method, no value is assigned. The correlation is obtained by the formula

$$r_g = 1 - \frac{6\Sigma\,(G)}{N^2 - 1}$$

For the data in Table 53, from column D,

$$G = 3+2+2+4+1+1+4+1+1.5+.5+1 = 21.0$$

and

$$r_g = .850$$

An approximation to r is given by the formula

$$r = sin\left(\frac{\pi}{2} r_g\right)$$

For $r_g = .850$, we have $\frac{180^\circ}{2} \times .850 = 76.50^\circ$. From tables giving trigonometric values the logarithm of the *sin* of an angle of 76.50° is 9.98843 −10. From a table of common logarithms the number corresponding to this logarithm is found to be .9737. It is clear that there is considerable deviation between r_g and r, and therefore this method should be used only as a means of obtaining a general idea as to whether or not any relationship exists. It is only a rough approximation and the value of r_g is not so readily comparable with r as is r_r, which is determined from ranks directly.

Coefficient of Contingency. There are some problems of relationship which it is desired to study where the nature of the material is such that it cannot be grouped readily into definite quantitative or numerical classes. For example, we may be interested in human traits, as the relation of eye color to hair color, or problems of this sort. There are also cases where one character may be recorded numerically while the character with which it is to be compared can be grouped only in broad categories. Then there may be other problems in which it is possible to measure both characters by some numerical methods but where for convenience it may be desirable to group the material in large groups and then determine the relation.

Pearson has given a measure to determine relationship in cases of this kind, and calls this constant the coefficient of contingency. For purposes of illustration we may use the data presented by Pearson, as given in Table 54, page 166. The first step is to determine what may be called independence frequencies. That is, on the assumption that there is no relation between the characters, the expected frequency for each observation is calculated.

TABLE 54

CALCULATION OF THE COEFFICIENT OF CONTINGENCY FOR THE RELATION IN ATHLETIC ABILITY BETWEEN BROTHERS IN THE SAME FAMILY

SECOND BROTHER	FIRST BROTHER			TOTAL
	ATHLETIC	BETWIXT	NON-ATHLETIC	
ATHLETIC	906 672.400	20 66.231	140 327.369	1066
BETWIXT	20 66.231	76 6.524	9 32.246	105
NON-ATHLETIC	140 327.369	9 32.246	370 159.385	519
TOTAL	1066	105	519	1690

$$\frac{\left(n_{rc}-\frac{n_r n_c}{N}\right)^2}{n_r n_c}$$

$$\begin{aligned}
(233.600)^2/1136356 &= .0480\\
(-46.231)^2/\ 111930 &= .0191\\
(-187.369)^2/\ 553254 &= .0635\\
(-46.231)^2/\ 111930 &= .0191\\
(69.476)^2/\ 11025 &= .4378\\
(-23.246)^2/\ 54495 &= .0099\\
(-187.369)^2/\ 553254 &= .0635\\
(-23.246)^2/\ 54495 &= .0099\\
(210.615)^2/\ 269361 &= .1647\\
\phi^2 &= .8355
\end{aligned}$$

The first number in each row and column is the observed value, and the number below it is the calculated independence frequency value. The columns to the right of the table are the steps in calculating ϕ^2.

$$\phi^2 = .8355$$

$$C_1 = \sqrt{\frac{\phi^2}{1+\phi^2}} = \sqrt{\frac{.8355}{1.8355}} = .675$$

This is done in the following way. The data in Table 54 show that for the athletic grouping according to the first brother there are 1066 individuals. The same is true for the same group for the second brother. It is noted that there are 906 cases where the first brother and second brother are both athletic. For this observed value the calculated value is determined by obtaining the product of 1066 × 1066, the totals of these athletic groupings, and dividing this product by the total number in the population. We have

$$\frac{1066\times1066}{1690}=672.400$$

This is the expected value on the assumption that there is perfect independence, or no correlation between the characters. The same method is used to obtain a calculated value for each of the groups or classes. Thus, for the case in the second column and first row where the observed number is 20, the calculated value is obtained from

$$\frac{105\times1066}{1690}$$

and is found to be 66.231.

After this has been done for all of the groups or classes in the table the differences between the observed and calculated values are obtained and squared. This squared value is divided by the product of the totals for the row and column concerned. For example, for the first observation, 906, the difference is 233.600. This is squared and divided by the product of the total for that row, 1066, and the total for that column, 1066, or 1136356. For the observed value 20 in the second row, the difference is 46.231. This difference is squared and divided by the product of the total of that row, 105, multiplied by the total of the column, 1066, and so on throughout the table. The sum of the several values is then obtained and is designated ϕ^2. The calculations that have just been completed may be represented by the following formula:

$$\textit{Let}\ \phi^2=\frac{\chi^2}{N}$$

$$\phi^2=\frac{\chi^2}{N}=\Sigma\frac{\left(n_{rc}-\frac{n_rn_c}{N}\right)^2}{n_r\,n_c}$$

which gives the mean square contingency. In this formula N refers to the number in the population; n_r to the number of individuals in any row and n_c to the number of individuals in any column; and n_{rc} to the number of individuals in any compartment common to both.

The formula that has been developed for the mean square contingency coefficient is

$$C_1=\sqrt{\frac{\phi^2}{1+\phi^2}}$$

The calculated value of ϕ^2 for the data in Table 54 is .8355. Substituting .8355 in the formula for C_1, the mean square contingency coefficient is .675.

Yule has suggested a method for obtaining the coefficient of contingency, and this method may be applied to the data in Table 54. The independence frequencies are obtained as before. Then the observed frequency is squared and divided by the independence frequency value, as illustrated in Table 55.

TABLE 55

VALUES OBTAINED IN CALCULATING COEFFICIENT OF CONTINGENCY FOLLOWING METHOD OF YULE

SQUARE OF OBSERVED VALUE DIVIDED BY CALCULATED VALUE	
$(906)^2/672.400$	1220.756
$(20)^2/66.231$	6.039
$(140)^2/327.369$	59.871
$(20)^2/66.231$	6.039
$(76)^2/6.524$	885.346
$(9)^2/32.246$	2.512
$(140)^2/327.369$	59.871
$(9)^2/32.246$	2.512
$(370)^2/159.385$	858.926
	$S=3101.872$

$$C_1=\sqrt{\frac{S-N}{S}}$$

$$C_1=\sqrt{\frac{3101.872-1690}{3101.872}}$$

$$=.675$$

For example, the first observed value, 906, is squared and divided by the independence value 672.400, giving 1220.756. This is done for each value in the table and the sum, 3101.872, is obtained. This sum is designated S. The coefficient of contingency is obtained from the formula

$$C_1=\sqrt{\frac{S-N}{S}}$$

where N is the number of individuals in the total population, in this particular case 1690. Substituting these values in the equation the coefficient of contingency is .675, which is the same as obtained by the first method.

Another illustration is given in Table 56, page 170, where the comparison is made between color of kernel and yield of wheat for different varieties.

The color character can be classified only in broad categories, as red and white. The yield, however, is expressed numerically and the several columns indicate how the individual strains have yielded as compared with a standard variety. Those varieties in the table to the left of the heavy vertical line yielded better than the standard variety, and those to the right of the line yielded less than the standard variety.

Using the same method as before the independence frequencies are obtained and are recorded in the table. From these ϕ^2 is found to be .343 and

$$C_1=\sqrt{\frac{.343}{1.343}}=.505$$

Using Yule's method S is found to be 479.763 and

$$C_1=\sqrt{\frac{479.763-357}{479.763}}=.506$$

For the types of examples illustrated here this method of measuring relationship is very useful. One objection is that the value varies somewhat with the method of grouping, but a correction

TABLE 56

THE RELATION BETWEEN COLOR OF KERNEL AND YIELD OF WHEAT, THE YIELD BEING EXPRESSED AS A GAIN OR LOSS WHEN COMPARED WITH A STANDARD TYPE

COLOR OF KERNEL	YIELD EXPRESSED AS A GAIN OR LOSS WHEN COMPARED WITH A STANDARD TYPE									TOTAL
	ABOVE 10.0	5.0 9.9	0 4.9	0 –4.9	–5.0 –9.9	–10.0 –14.9	–15.0 –19.9	–20.0 –24.9	ABOVE –25.0	
RED	4	8	33	17	24	9	1	3	0	99
	1.109	2.218	11.092	16.084	27.176	27.454	10.815	2.218	.832	
WHITE	0	0	7	41	74	90	38	5	3	258
	2.891	5.782	28.908	41.916	70.824	71.546	28.185	5.782	2.168	
TOTAL	4	8	40	58	98	99	39	8	3	357

may be made for this. Let R represent the number of rows and C the number of columns. By obtaining the value

$$\frac{(R-1)(C-1)}{N}$$

and subtracting it from ϕ^2 we have the corrected value for ϕ^2, and from this the corrected value of C_1 is obtained.

Thus, for the first problem above, the number of rows and the number of columns are the same, 3. By substituting these in

$$\frac{(R-1)(C-1)}{N}$$

we have

$$\frac{(3-1)(3-1)}{1690}=.0024$$

Subtracting this value from ϕ^2 we have

$$.8355-.0024=.8331$$

and with this corrected value for ϕ^2

$$C_1=\sqrt{\frac{.8331}{1.8331}}=.674$$

There is only a slight change in the value of C_1 in the third decimal place.

Applying the correction to the second example, where $R=2$ and $C=9$, we have

$$\frac{(R-1)(C-1)}{N}=\frac{(2-1)(9-1)}{357}=.022$$

Correcting ϕ^2 we have .493 as the corrected value for C_1. In this case the value of the coefficient of contingency is reduced by the amount of .012.

The value of C_1 with restrictions as to grouping and where the distribution is normal or practically so, approaches the correlation coefficient, *r*. It is usually better, however, to apply the method of contingency to a five by five grouping since the coefficient is not so reliable when less than about 16 compartments or cells are used for the calculation. Therefore the method is strictly applicable only to large samples, and when used for examples with fewer than 16 cells care must be exercised in interpretation.

CHAPTER VIII

MULTIPLE AND PARTIAL CORRELATION

In discussing the measurement of correlation thus far we have considered the relationship between two variables only. It happens, however, that there are numerous problems in which it is desired to know how several variables may act together to affect one variable. For example, if we are interested in predicting the yield of a crop it is necessary to know more than the effect of rainfall or the amount of sunshine on the crop, or of any other single factor that may influence the crop. In addition to the effect of each of the characters alone it is necessary for accurate prediction to determine the result of all of the environmental factors together on the crop yield. The methods of correlation analysis have been extended to make such determinations possible, and lead to the field of multiple correlation.

Multiple Correlation. By multiple correlation is meant the measurement of the effect of several variables on one variable. As an illustration we may use the data in Table 57, page 174, giving the weight per bushel, weight per 100 kernels, weight of straw, and yield of grain for 25 varieties of oats. For convenience the characters are designated by A, B, C, and Y, respectively, as indicated in the table. We will discuss in order the effect of one, two, and three variables on the yield of grain, and will consider to what extent yield of grain is dependent on the other three characters.

We will proceed to determine the effect of these characters on the yield, and will approach the problem by determining first the simple correlation between yield of grain, Y, and weight per bushel, A. This correlation will be determined without grouping from the following formula:

$$r_{ya} = \frac{\Sigma AY - (\Sigma A)\,M_y}{\sqrt{\Sigma A^2 - (\Sigma A)\,M_a}\,\sqrt{\Sigma Y^2 - (\Sigma Y)\,M_y}}$$

TABLE 57

DATA ON 25 VARIETIES OF OATS, USED TO ILLUSTRATE ANALYSIS OF MULTIPLE CORRELATION

OBSERVATION NUMBER	WEIGHT PER BUSHEL, POUNDS A	WEIGHT PER 100 KERNELS, GRAMS B	WEIGHT OF STRAW, TONS C	YIELD OF GRAIN, BUSHELS Y
1	30.50	2.234	.87	47.5
2	30.80	2.074	1.00	55.5
3	31.62	2.250	1.57	60.4
4	31.81	2.420	1.49	61.9
5	31.50	2.584	1.60	70.2
6	31.81	2.518	1.56	66.8
7	34.81	2.492	1.37	62.4
8	31.75	2.774	1.38	56.6
9	33.69	2.616	1.42	68.0
10	32.62	2.700	1.44	67.6
11	31.94	2.764	1.37	53.8
12	32.75	2.760	1.42	60.5
13	32.00	2.644	1.51	63.6
14	31.50	2.734	1.63	67.3
15	31.94	2.710	1.40	68.1
16	31.56	2.274	1.53	65.4
17	32.50	2.850	1.44	66.8
18	32.44	2.824	1.52	65.1
19	34.00	2.584	1.44	71.3
20	32.75	2.614	1.50	73.9
21	34.69	2.830	1.41	70.2
22	33.06	2.844	1.39	69.5
23	29.75	2.230	1.02	50.9
24	30.37	2.066	.87	46.3
25	33.00	2.844	1.32	66.1

By substitution of the proper symbols in this formula (shown on bottom of the page 173) the correlation coefficients for the several variables may be obtained. For example,

$$r_{ab}=\frac{\Sigma AB-(\Sigma A)\,M_b}{\sqrt{\Sigma A^2-(\Sigma A)\,M_a}\;\sqrt{\Sigma B^2-(\Sigma B)\,M_b}}$$

The meaning of these symbols and the values needed for solving for the several correlation coefficients are given in Table 58, page 175. The standard deviations may be obtained by dividing the last value in each column by $\sqrt{N-1}$.

TABLE 58
SYMBOLS AND VALUES NEEDED FOR DETERMINING SIMPLE CORRELATION COEFFICIENTS

	A	B	C	Y
Number of Individuals	$N = 25$			
Sum of Values	$\Sigma A = 805.16$	$\Sigma B = 64.234$	$\Sigma C = 34.47$	$\Sigma Y = 1575.7$
Mean	$M_a = 32.206$	$M_b = 2.569$	$M_c = 1.379$	$M_y = 63.028$
Sum × Mean	$(\Sigma A)M_a = 25930.98$	$(\Sigma B)M_b = 165.02$	$(\Sigma C)M_c = 47.53$	$(\Sigma Y)M_y = 99313.22$
Sum × Mean of Correlated Variable	$(\Sigma A)M_y = 50747.62$ $(\Sigma A)M_b = 2068.46$	$(\Sigma B)M_y = 4048.54$ $(\Sigma B)M_c = 88.58$	$(\Sigma C)M_y = 2172.58$	
	$(\Sigma A)M_c = 1110.32$			
Sum of Products of Correlated Variables	$\Sigma AY = 50894.36$ $\Sigma AB = 2073.21$	$\Sigma BY = 4075.67$ $\Sigma BC = 89.27$	$\Sigma CY = 2202.36$	
	$\Sigma AC = 1113.00$			
Sum of Squares of Values	$\Sigma A^2 = 25968.68$	$\Sigma B^2 = 166.53$	$\Sigma C^2 = 48.60$	$\Sigma Y^2 = 100648.49$
Sum of Squares — (Sum of Values × Mean)	$\Sigma A^2 - (\Sigma A)M_a = 37.70$	$\Sigma B^2 - (\Sigma B)M_b = 1.51$	$\Sigma C^2 - (\Sigma C)M_c = 1.07$	$\Sigma Y^2 - (\Sigma Y)M_y = 1335.27$
Square Root	$\sqrt{\Sigma A^2 - (\Sigma A)M_a} = 6.14$	$\sqrt{\Sigma B^2 - (\Sigma B)M_b} = 1.23$	$\sqrt{\Sigma C^2 - (\Sigma C)M_c} = 1.03$	$\sqrt{\Sigma Y^2 - (\Sigma Y)M_y} = 36.54$

Simple correlation coefficients determined from above values

$r_{ay} = .654$
$r_{ab} = .629$
$r_{ac} = .424$
$r_{by} = .604$
$r_{bc} = .543$
$r_{cy} = .791$

In the discussion of partial correlation coefficients, A, B, C, and Y are the same as 1, 2, 3, and 4.

Substituting from Table 58 the necessary values in the formula for the simple correlation of Y and A we have

$$r_{ya}=\frac{50894.36-50747.62}{6.14\times 36.54}=\frac{146.74}{224.36}=.654$$

The simple correlation coefficients for the other variables have been obtained by the same process, and the several correlation coefficients are recorded in Table 58.

The value for r_{ya} (.654) is the simple correlation between the two variables, and it indicates that yield of grain is dependent to some extent on the weight per bushel. In Chapter VI we learned how to use the correlation coefficient to form a regression equation for the purpose of prediction. We may now use the correlation coefficient just obtained to predict the yield of grain on the basis of our knowledge of the weight per bushel. For this prediction we have the following regression equation, in which Y_a is the estimated or predicted value of Y based on the observed value of A:

$$Y_a=M_y+r\,\frac{\sqrt{\Sigma Y^2-(\Sigma Y)\,M_y}}{\sqrt{\Sigma A^2-(\Sigma A)\,M_a}}\,(A-M_a)$$

Substituting in this equation the numerical values from Table 58, we have

$$\begin{aligned}Y_a&=63.028+.654\,\frac{36.54}{6.14}\,(A-32.206)=63.028+(.654\times 5.951)\,(A-32.206)\\&=63.028+3.892\,(A-32.206)=63.028+3.892A-125.346\\Y_a&=3.892A-62.318\end{aligned}$$

To obtain the value of Y_a for the first observed value of A, 30.50 from Table 57 we have

$$Y_a=(3.892\times 30.50)-62.318=56.388$$

Substituting the other observed values for A we obtain the predicted yields, *P*, of Y as given in column 2 of Table 59, page 177. For convenience the predicted yields in this table are read to one decimal only.

These predicted yields may be compared with the actual yields, giving the errors of estimate as recorded in column 3 of Table 59.

TABLE 59

OBSERVED YIELDS OF GRAIN, WITH YIELDS AND ERRORS OF ESTIMATE PREDICTED FROM ONE, TWO, AND THREE VARIABLES

OBSERVED YIELD OF GRAIN *O* 1	PREDICTED FROM WEIGHT PER BUSHEL (A) *P* 2	ERRORS OF ESTIMATE *O–P* 3	PREDICTED FROM WEIGHT PER BUSHEL (A) AND WEIGHT PER 100 KERNELS (B) *P* 4	ERRORS OF ESTIMATE *O–P* 5	PREDICTED FROM WEIGHT PER BUSHEL (A), WEIGHT PER 100 KERNELS (B), AND WEIGHT OF STRAW (C) *P* 6	ERRORS OF ESTIMATE *O–P* 7
47.5	56.4	–8.9	55.2	–7.7	47.8	– .3
55.5	57.6	–2.1	54.5	1.0	51.1	4.4
60.4	60.7	– .3	58.4	2.0	65.5	–5.1
61.9	61.5	.4	60.5	1.4	64.4	–2.5
70.2	60.3	9.9	61.3	8.9	66.3	3.9
66.8	61.5	5.3	61.5	5.3	66.0	.8
62.4	73.2	–10.8	69.3	–6.9	68.5	–6.1
56.6	61.3	–4.7	63.7	–7.1	62.3	–5.7
68.0	68.8	– .8	67.5	.5	67.2	.8
67.6	64.6	3.0	65.4	2.2	65.4	2.2
53.8	62.0	–8.2	64.2	–10.4	62.5	–8.7
60.5	65.1	–4.6	66.3	–5.8	65.3	–4.8
63.6	62.2	1.4	63.2	.4	65.5	–1.9
67.3	60.3	7.0	62.7	4.6	67.1	.2
68.1	62.0	6.1	63.6	4.5	63.1	5.0
65.4	60.5	4.9	58.5	6.9	64.5	.9
66.8	64.2	2.6	66.5	.3	65.3	1.5
65.1	63.9	1.2	66.1	–1.0	66.9	–1.8
71.3	70.0	1.3	68.0	3.3	68.3	3.0
73.9	65.1	8.8	64.9	9.0	66.9	7.0
70.2	72.7	–2.5	72.2	–2.0	69.5	.7
69.5	66.4	3.1	67.9	1.6	65.5	4.0
50.9	53.5	–2.6	53.2	–2.3	49.4	1.5
46.3	55.9	–9.6	53.3	–7.0	47.3	–1.0
66.1	66.1	0.0	67.8	–1.7	63.8	2.3

Some of these differences are positive and some are negative, and if a sufficient number of decimals had been retained the sum of these errors of estimate would equal zero.

In the equation that has been used to obtain the predicted yields the value 3.892 may be considered as the regression coefficient, and it indicates the amount of difference we may expect to find in the yield of grain with a difference of one unit in weight per bushel For example, variety 1 has a weight per bushel of 30.50 and its predicted yield, read to three decimals, is 56.388. Variety 5 has a weight per bushel of 31.50, one unit larger, and its predicted yield is 60.280. This is larger than the predicted yield of variety 1 by 60.280–56.388, or 3.892, which is the value of the regression coefficient.

As usual, there are two regression equations for each correlation. The regression equation just determined gives the predicted values of Y from observed values of A. We may obtain the predicted values of A from observed values of Y from the following equation, in which A_y means the predicted weight per bushel of A from observed yields of Y:

$$A_y = M_a + r \frac{\sqrt{\Sigma A^2 - (\Sigma A)\ M_a}}{\sqrt{\Sigma Y^2 - (\Sigma Y)\ M_y}} (Y - M_y)$$

Substituting in this equation the numerical values from Table 58 we have

$$A_y = 32.206 + .654 \frac{6.14}{36.54} (Y - 63.028) = 32.206 + (.654 \times .168)\ (Y - 63.028)$$

$$= 32.206 + .110\ (Y - 63.028) = 32.206 + .110Y - 6.933$$

$$A_y = .110Y + 25.273$$

Here the value .110 may be considered the regression coefficient, and indicates the amount of difference that may be expected in the weight per bushel with a difference of one unit in the yield of grain. It is unnecessary here to calculate the predicted values for A based on the observed yields of Y.

It may be of interest to show the relation between the correlation coefficient, *r*, and these two regression coefficients, 3.892 and

.110. This is done by taking the square root of the product of these two values, as

$$r_{ya} = \sqrt{3.892 \times .110} = .654$$

In other words, the correlation coefficient is really the geometric mean of the two regression coefficients. This result may be used as a check on the calculations of the regression coefficients.

Since from column 3 of Table 59 there is apparent a certain amount of variation between the predicted and the observed values of Y, it is desirable to determine the standard deviation of the predicted values, as was done in Chapter VI. This standard deviation may be obtained as follows:

$$\sigma_{y \cdot a} = \sqrt{\frac{(1-r^2)\,[\Sigma Y^2 - (\Sigma Y)\,M_y]}{N-2}}$$

In this formula $N-2$ is used in place of N, and in following formulas $N-3$ and $N-4$ will be used in place of N. The reason for this will be explained in Chapter XIII.

Substituting in this formula the numerical values from Table 58 we have

$$\sigma_{y \cdot a} = \sqrt{\frac{[1-(.654)^2]\,(1335.27)}{25-2}} = \sqrt{\frac{.572 \times 1335.27}{23}} = \sqrt{\frac{763.774}{23}}$$
$$= \sqrt{33.207565} = 5.763$$

We find the standard deviation of the estimate to be 5.763. This standard deviation of estimate may also be determined by squaring the errors of estimate in column 3 of Table 59, summing these squares and dividing by $N-2$, and extracting the square root of this quotient, but the values are more readily obtained by use of the formula for the standard deviation of estimate of predicted values. The values obtained by the two methods will agree exactly only when a sufficient number of decimals are retained.

The standard deviation of estimate measured on the basis of regression, 5.763, is lower than the standard deviation of Y, 7.459, determined from the mean. This leads to the conclusion that the variation of the weight per bushel accounts for some of the variability in yield of grain. By obtaining the value

$$\frac{100\times 5.763}{7.459}=77.26$$

it is seen that the new standard deviation is 77.26 per cent of the original standard deviation, and considering the original standard deviation as 100 per cent we find that there is a reduction in the average of variability of 22.74 per cent by using the information available on weight per bushel.

The percentage of reduction that may be expected from the standard deviation obtained from the mean to the standard deviation of estimate may also be obtained from the following general formula. The percentage reduction from σ_y to $\sigma_{y.a}$ equals

$$100\left[1-\sqrt{\frac{N-1}{N-2}\,(1-r^2)}\,\right]$$

Using the value of r_{ya} (.654) obtained for the 25 individuals, we have

$$100\left[1-\sqrt{\frac{25-1}{25-2}\,[1-(.654)^2]}\,\right]=22.72$$

This percentage reduction differs slightly from that obtained above, due to the number of decimals retained. It must be clear that for a large number of individuals the factor $\frac{N-1}{N-2}$ will have very little effect. For low correlation and a small sample the effect is appreciable.

The amount of reduction that may be expected from different values of r for $N=100$ is given in Table 60, page 181, taken from Wallace and Snedecor. This table gives the expected reduction in the standard deviation calculated on the basis of the regression.

TABLE 60

PERCENTAGE OF REDUCTION EXPECTED FROM STANDARD DEVIATION OBTAINED FROM THE MEAN TO STANDARD DEVIATION OF ESTIMATE

$N=100$

r	PERCENTAGE REDUCTION	r	PERCENTAGE REDUCTION
.30	4.1	.92	60.6
.40	7.9	.94	65.7
.50	13.0	.95	68.6
.55	16.1	.96	71.9
.60	19.6	.97	75.6
.65	23.6	.98	80.0
.70	28.2	.99	85.8
.75	33.5	.995	90.0
.80	39.7	.999	95.5
.85	47.1	.9995	96.8
.90	56.2	.9999	98.6

We may now proceed to determine the effect of two variables, weight per bushel (A) and weight per 100 kernels (B), on the yield of grain (Y), or to predict the values of Y from observed values of A and B. To do this we make use of the correlation between Y and A and between Y and B, and then analyze the effect of A and B together on Y. From Table 58 we have the necessary correlation coefficients

$$r_{ay}=.654$$
$$r_{by}=.604$$
$$r_{ab}=.629$$

It may be pointed out that r_{ay} is the same as r_{ya}, r_{by} is the same as r_{yb}, r_{ab} is the same as r_{ba}, and so on, and the order of writing these subscripts is unimportant.

From this point the procedure differs somewhat from that when considering the relationship between two variables, and it is desirable now to make use of what may be called the standard partial regression coefficients, which may be designated β_{ya} and β_{yb}. These may be obtained from the following equations:

$$\text{Equation I} \quad \beta_{ya} \quad + r_{ab}\beta_{yb} = r_{ya}$$
$$\text{Equation II} \quad r_{ab}\beta_{ya} + \beta_{yb} \quad = r_{yb}$$

From these equations by solving for β_{ya} we have

$$\text{Equation I} = \beta_{ya} + r_{ab}\beta_{yb} = r_{ay}$$

Multiplying Equation II
by r_{ab}

$$r^2_{ab}\beta_{ya} + r_{ab}\beta_{yb} = r_{ab}r_{yb}$$

Subtracting

$$\beta_{ya} - r^2_{ab}\beta_{ya} = r_{ya} - (r_{ab}r_{yb})$$

Factoring

$$\beta_{ya}(1 - r^2_{ab}) = r_{ya} - (r_{ab}r_{yb})$$

Dividing by $(1 - r^2_{ab})$

$$\beta_{ya} = \frac{r_{ya} - (r_{ab}\,{}_{yb})}{1 - r^2_{ab}}$$

By a similar process

$$\beta_{yb} = \frac{r_{yb} - (r_{ab}r_{ya})}{1 - r^2_{ab}}$$

Substituting the known values, that is the correlation coefficients r_{ya}, r_{yb}, and r_{ab}, we have

$$\beta_{ya} = \frac{.654 - (.629 \times .604)}{1 - (.629)^2} = .454$$

By substituting the correlation coefficients in the formula for β_{yb} we have

$$\beta_{yb} = .320$$

The two β values may also be found by substituting directly the proper r values in the general equations, as

$$\text{Equation I} \quad 1.000\beta_{ya} + .629\beta_{yb} = .654$$
$$\text{Equation II} \quad .629\beta_{ya} + 1.000\beta_{yb} = .604$$

Solving in the usual way we find

$$\beta_{ya} = .454$$
$$\beta_{yb} = .320$$

Using these β values we may now obtain the regression equation for estimating Y values from the observed values of the two variables, A and B. Up to this point we have used regression equations based on one variable only, but now it becomes necessary to form a new regression equation based on two variables. This gives what is known as a multiple regression equation and the formula is

$$Y_{ab}=M_y+\beta_{ya}\frac{\sqrt{\Sigma Y^2-(\Sigma Y)\ M_y}}{\sqrt{\Sigma A^2-(\Sigma A)\ M_a}}(A-M_a)+\beta_{yb}\frac{\sqrt{\Sigma Y^2-(\Sigma Y)\ M_y}}{\sqrt{\Sigma B^2-(\Sigma B)\ M_b}}(B-M_b)$$

In this equation Y_{ab} represents the predicted values of Y from the observed values of both A and B.

Substituting the known values we have

$$\begin{aligned}
Y_{ab}&=63.028+.454\ \frac{36.54}{6.14}\ (A-32.206)+.320\ \frac{36.54}{1.23}\ (B-2.569)\\
&=63.028+(.454\times 5.951)\ (A-32.206)+(.320\times 29.707)\ (B-2.569)\\
&=63.028+2.702\ (A-32.206)+9.506\ (B-2.569)\\
&=63.028+2.702A-87.021+9.506B-24.421\\
Y_{ab}&=2.702A+9.506B-48.414
\end{aligned}$$

For determining the predicted value for Y based on the observed values of A and B for the first variety in Table 57, we have

$$Y_{ab}=2.702\ (30.50)+9.506\ (2.234)-48.414=55.2$$

Applying this multiple regression equation and using the data in Table 57 we obtain the values in column 4 of Table 59. These are the predicted yields of grain when considering both weight per bushel and weight per 100 kernels.

At this point we may consider the multiple correlation coefficient. The regression equation indicates how the predicted value of one variable may be obtained from the observed values of two other variables, but it is also important to have some measure that will indicate this relationship. Such a measure is the multiple correlation coefficient. We may use the values determined for β_{ya} and β_{yb}, r_{ya} and r_{ab} to obtain the multiple correlation coefficient from the formula

$$R^2=r_{ya}\beta_{ya}+r_{yb}\beta_{yb}$$

Substituting the numerical values we have

$$\begin{aligned} R^2 &= (.654\times .454)+(.604\times .320) \\ &= .296916+.193280=.490196 \\ R &= \sqrt{.490196}=.700 \end{aligned}$$

We may now consider the meaning of this multiple correlation coefficient. This coefficient, $R=.700$, is the value of the simple correlation between the actual yields and predicted yields as given in columns 1 and 4 of Table 59. The correlation between yield of grain and weight per bushel is .654, while the coefficient of multiple correlation when the effect of weight per 100 kernels is added is .700. Thus there is some slight improvement on the estimation when two variables are considered.

The errors of estimate for the effect of two variables are given in column 5 of Table 59. It is seen that there is some slight reduction in the errors of estimate when the effect of two variables is considered as compared with the effect of only one variable (column 3).

The standard deviation of the estimate may be obtained from the formula

$$\sigma_{y.ab}=\sqrt{\frac{[1-R^2]\,[\Sigma Y^2-(\Sigma Y)\,M_y]}{N-3}}$$

Substituting the numerical values the standard deviation of estimate is found to be 5.564. Making use of this standard deviation to determine the amount of reduction due to the effect of two variables, we find

$$\frac{100\times 5.564}{7.459}=74.59$$

Compared with the standard deviation obtained from the mean we have, by considering the effect of two variables, A and B, reduced the standard deviation by 25.41 per cent.

It is often important to consider the effect of several variables together, and this may be done by determining additional standard partial regression coefficients obtained from supplemental equations. These equations may be extended by adding a term to each equation and an additional equation for each additional variable. For the effect of three variables on a fourth the equations are

$$\begin{aligned}
\beta_{ya}+r_{ab}\beta_{yb}+r_{ac}\beta_{yc}&=r_{ya}\\
r_{ba}\beta_{ya}+\beta_{yb}+r_{bc}\beta_{yc}&=r_{yb}\\
r_{ca}\beta_{ya}+r_{cb}\beta_{yb}+\beta_{yc}&=r_{yc}
\end{aligned}$$

For the effect of four variables on a fifth the equations are

$$\begin{aligned}
\beta_{ya}+r_{ab}\beta_{yb}+r_{ac}\beta_{yc}+r_{ad}\beta_{yd}&=r_{ya}\\
r_{ba}\beta_{ya}+\beta_{yb}+r_{bc}\beta_{yc}+r_{bd}\beta_{yd}&=r_{yb}\\
r_{ca}\beta_{ya}+r_{cb}\beta_{yb}+\beta_{yc}+r_{cd}\beta_{yd}&=r_{yc}\\
r_{da}\beta_{ya}+r_{db}\beta_{yb}+r_{dc}\beta_{yc}+\beta_{yd}&=r_{yd}
\end{aligned}$$

For the effect of five variables on a sixth the equations are

$$\begin{aligned}
\beta_{ya}+r_{ab}\beta_{yb}+r_{ac}\beta_{yc}+r_{ad}\beta_{yd}+r_{ae}\beta_{ye}&=r_{ya}\\
r_{ba}\beta_{ya}+\beta_{yb}+r_{bc}\beta_{yc}+r_{bd}\beta_{yd}+r_{be}\beta_{ye}&=r_{yb}\\
r_{ca}\beta_{ya}+r_{cb}\beta_{yb}+\beta_{yc}+r_{cd}\beta_{yd}+r_{ce}\beta_{ye}&=r_{yc}\\
r_{da}\beta_{ya}+r_{db}\beta_{yb}+r_{dc}\beta_{yc}+\beta_{yd}+r_{de}\beta_{ye}&=r_{yd}\\
r_{ea}\beta_{ya}+r_{eb}\beta_{yb}+r_{ec}\beta_{yc}+r_{ed}\beta_{yd}+\beta_{ye}&=r_{ye}
\end{aligned}$$

For the effect of six variables on a seventh the equations are

$$\begin{aligned}
\beta_{ya}+r_{ab}\beta_{yb}+r_{ac}\beta_{yc}+r_{ad}\beta_{yd}+r_{ae}\beta_{ye}+r_{af}\beta_{yf}&=r_{ya}\\
r_{ba}\beta_{ya}+\beta_{yb}+r_{bc}\beta_{yc}+r_{bd}\beta_{yd}+r_{be}\beta_{ye}+r_{bf}\beta_{yf}&=r_{yb}\\
r_{ca}\beta_{ya}+r_{cb}\beta_{yb}+\beta_{yc}+r_{cd}\beta_{yd}+r_{ce}\beta_{ye}+r_{cf}\beta_{yf}&=r_{yc}\\
r_{da}\beta_{ya}+r_{db}\beta_{yb}+r_{dc}\beta_{yc}+\beta_{yd}+r_{de}\beta_{ye}+r_{df}\beta_{yf}&=r_{yd}\\
r_{ea}\beta_{ya}+r_{eb}\beta_{yb}+r_{ec}\beta_{yc}+r_{ed}\beta_{yd}+\beta_{ye}+r_{ef}\beta_{yf}&=r_{ye}\\
r_{fa}\beta_{ya}+r_{fb}\beta_{yb}+r_{fc}\beta_{yc}+r_{fd}\beta_{yd}+r_{fe}\beta_{ye}+\beta_{yf}&=r_{yf}
\end{aligned}$$

It is possible to extend these equations for any number of variables, but the work of solving them becomes laborious and a more convenient method will be presented later.

For the effect of the three variables A, B, and C on the fourth variable Y, we may substitute in the first set of equations above the correlation coefficients from Table 58.

EQUATION I $1.000\ \beta_{ya}+\ .629\ \beta_{yb}+\ .424\ \beta_{yc}=.654$

EQUATION II $.629\ \beta_{ya}+1.000\ \beta_{yb}+\ .543\ \beta_{yc}=.604$

EQUATION III $.424\ \beta_{ya}+\ .543\ \beta_{yb}+1.000\ \beta_{yc}=.791$

These equations are solved in the usual way. Rewriting Equation II and multiplying Equation I by .629, we have

$$\begin{array}{llcl} \text{EQUATION II} & .629\ \beta_{ya}+1.000\ \beta_{yb}+.543\ \beta_{yc} & = & .604 \\ \text{EQUATION I} & .629\ \beta_{ya}+\ .396\ \beta_{yb}+.267\ \beta_{yc} & = & .411 \end{array}$$

Subtracting

$$.604\ \beta_{yb}+.276\ \beta_{yc}=.193 \qquad (1)$$

Rewriting Equation III and multiplying Equation I by .424, we have

$$\begin{array}{llcl} \text{EQUATION III} & .424\ \beta_{ya}+.543\ \beta_{yb}+1.000\ \beta_{yc} & = & .791 \\ \text{EQUATION I} & .424\ \beta_{ya}+.267\ \beta_{yb}+\ .180\ \beta_{yc} & = & .277 \end{array}$$

Subtracting

$$.276\ \beta_{yb}+\ .820\ \beta_{yc}=.514 \qquad (2)$$

Multiplying (2) by .604 and (1) by .276, we have

$$\begin{array}{llcl} (2) & .167\ \beta_{yb}+.495\ \beta_{yc} & = & .310 \\ (1) & .167\ \beta_{yb}+.076\ \beta_{yc} & = & .053 \end{array}$$

Subtracting

$$\begin{array}{rcl} .419\ \beta_{yc} & = & .257 \\ \beta_{yc} & = & .613 \end{array}$$

Substituting the value for β_{yc} in (1)

$$\begin{array}{lrcl} (1) & .604\ \beta_{yb}+.169 & = & .193 \\ & .604\ \beta_{yb} & = & .024 \\ & \beta_{yb} & = & .040 \end{array}$$

Substituting the values of β_{yb} and β_{yc} in Equation I

$$\begin{array}{ll} \text{EQUATION I} & 1.000\ \beta_{ya}+(.629\times.040)+(.424\times.613)=.654 \\ & \beta_{ya}=.369 \end{array}$$

Using these β values we may form the regression equation for predicting the yield of grain, Y, from the observed values of A, B, and C. By following the method for the regression equation for two variables, and adding a third, the regression equation may be written as follows:

$$Y_{abc}=M_y+\beta_{ya}\frac{\sqrt{\Sigma Y^2-(\Sigma Y)M_y}}{\sqrt{\Sigma A^2-(\Sigma A)M_a}}\ (A-M_a)$$

$$+\beta_{yb}\frac{\sqrt{\Sigma Y^2-(\Sigma Y)M_y}}{\sqrt{\Sigma B^2-(\Sigma B)M_b}}\ (B-M_b)$$

$$+\beta_{yc}\frac{\sqrt{\Sigma Y^2-(\Sigma Y)M_y}}{\sqrt{\Sigma C^2-(\Sigma C)M_c}}\ (C-M_c)$$

Substituting the numerical values from Table 58 and the necessary β values in this equation we have

$$Y_{abc}=63.028+.369\frac{36.54}{6.14}\ (A-32.206)$$

$$+.040\frac{36.54}{1.23}\ (B-\ 2.569)$$

$$+.613\frac{36.54}{1.03}\ (C-\ 1.379)$$

Completing the necessary calculations we have

$$Y_{abc}=3.196A+1.188B+21.747C-40.737$$

Substituting the several values for A, B, and C for the individual varieties given in Table 57, we obtain the predicted values as given in the sixth column of Table 59. For example, for the predicted value of Y based on the observed values of A, B, and C for the first variety in Table 57, we have

$$Y_{abc}=3.196(30.50)+1.188(2.234)+21.747(.87)-40.737=47.8$$

It is noted that the predicted values agree more closely with the observed values when we use our knowledge of weight per bushel, weight per 100 kernels, and weight of straw. This is evident from the last column in Table 59, giving the errors of estimate.

Following a method similar to that for two variables, we may now determine the coefficient of multiple correlation from the equation

$$R^2=\beta_{ya}r_{ay}+\beta_{yb}r_{by}+\beta_{yc}r_{cy}$$

Substituting in this equation the numerical values obtained we have

$$R^2=(.369)(.654)+(.040)(.604)+(.613)(.791)=.750369$$

$$R=\sqrt{.750369}=.866$$

This gives a higher value for R than was obtained with the two variables, weight per bushel and weight per 100 kernels, which was .700. Again it may be well to point out that this value, $R = .866$, is the simple correlation between the predicted and actual yields of columns 1 and 6 of Table 59. It is seen that a gradual improvement has been made by adding more variables in our study. The correlation coefficient between yield and one variable is .654, between yield and two variables it is .700, and when the effect of a third variable is added it becomes .866.

In order to determine the amount of reduction that has been made by the addition of the variable, weight of straw, we determine the standard deviation of estimate from the formula

$$\sigma_{y \cdot abc} = \sqrt{\frac{[1 - R^2][\Sigma Y^2 - (\Sigma Y)M_y]}{N - 4}}$$

Substituting the numerical values in this formula we find

$$\sigma_{y \cdot abc} = 3.987$$

To determine from this standard deviation of estimate the amount of reduction we have

$$\frac{100 \times 3.987}{7.459} = 53.45$$

This shows that there has been a reduction of 46.55 per cent in the standard deviation when based on the regression of these three variables, as compared with the standard deviation of the yield measured from the mean. We may conclude that in this case yield depends considerably on the effect of the three variables studied. The extent of this dependence may be determined from the following general formula

$$100 \times 1 - \sqrt{1 - R^2}$$

Substituting the coefficient of multiple correlation we have

$$100 \times 1 - \sqrt{1 - (.866)^2} = 50.00$$

TABLE 61

DATA ON 25 OAT PLANTS, USED TO ILLUSTRATE ANALYSIS OF MULTIPLE CORRELATION FOR SEVERAL VARIABLES

OBSERVATION NUMBER	AVERAGE HEIGHT OF PLANT, CENTIMETERS A	AVERAGE NUMBER OF SPIKELETS PER CULM B	NUMBER OF CULMS C	AVERAGE NUMBER OF KERNELS PER CULM D	TOTAL YIELD OF GRAIN, GRAMS Y	*S*
1	50	20	2	25	1.0	98.0
2	59	20	2	45	2.5	128.5
3	63	24	3	43	2.5	135.5
4	66	30	4	42	3.0	145.0
5	52	18	2	28	1.5	101.5
6	69	40	5	40	4.0	158.0
7	71	40	6	30	4.0	151.0
8	74	40	5	60	5.2	184.2
9	78	44	4	75	6.0	207.0
10	55	24	3	34	2.0	118.0
11	84	51	4	90	7.0	236.0
12	60	25	3	37	2.0	127.0
13	80	50	6	52	6.0	194.0
14	64	30	4	40	3.0	141.0
15	70	35	4	45	4.2	158.2
16	79	45	5	68	6.5	203.5
17	77	50	6	50	5.0	188.0
18	65	35	5	40	3.2	148.2
19	67	32	3	50	3.5	155.5
20	70	38	5	40	4.5	157.5
21	84	50	5	80	7.2	226.2
22	76	45	4	65	4.8	194.8
23	73	44	6	42	5.0	170.0
24	72	36	4	45	4.5	161.5
25	75	43	6	35	4.0	163.0
TOTAL	1733	909	106	1201	102.1	4051.1

This shows that the three variables, weight per bushel, weight per 100 kernels, and weight of straw, account for 50.00 per cent of variability of the yield, leaving 50.00 per cent unaccounted for, or in other words 50.00 per cent is dependent on characters that have not been analyzed in this study.

The foregoing illustrates how the method of multiple correlation may be used to determine the effect of several characters together on one particular character. If more characters for these oat varieties were available it would be possible to include them and determine the effect of all the characters on the yield. There are many other problems where multiple correlation may be very useful. For example, in the field of economics in the study of farm incomes, the size of farm, the number of laborers, the number of animals, the yields of different crops, and the like, may be studied in relation to the farmer's income.

When it is desired to determine the multiple correlation with several variables under consideration, it is better to approach the solution in a different way. Using the data in Table 61 we may illustrate a more convenient way for determining multiple correlation and the regression coefficients, or β values, for several variables than can be done by solving the necessary simultaneous equations.

In Table 61, page 189, the characters have been given the symbols A, B, C, D, and Y. It will be noted that the last column in the table has been designated S. This column is added for the purpose of simplifying the checking of calculations to be made later, and the values in this column are obtained by summing the numbers in the other columns in turn. Thus, for the first individual we have

$$50+20+2+25+1.0=98.0$$

Proceeding in the same way the other values in column S are obtained. Column S is treated as another variable, and will appear in most of the calculations.

For determining the multiple correlation and the regression coefficients for the several variables we will follow the method suggested by Wallace and Snedecor, and to obtain the necessary correlation coefficients and β values conveniently Tables 62, 63, 64, and 65 have been arranged in systematic order. Tables 62 (pages 192–3) and 63 (page 194) cover the steps leading to the determination of the correlation coefficients, and Tables 64 (page 196) and 65 (page 197) proceed to the calculation of the β values. In Tables 62 and 64 the necessary operations are indicated by symbols, and in Tables 63 and 65 the numerical values have been entered. These tables may serve as guides and to them may be added the columns and lines for any desired number of variables.

The meaning of the various symbols and the calculations should be clear. The different characters are indicated by the letters as at the head of the columns in Table 61. As usual, M refers to the mean for the variable indicated by the subscript, as for example M_a is the mean of A, and similarly for the other variables. The numerical values as determined from the data in Table 61 are recorded in Table 63 in accordance with the symbols given in Table 62.

The use of column S in Table 63 for purposes of checking the calculations may now be explained. The first number in column S is the sum of the numbers in the preceding columns. The means for the several variables have been obtained and the mean for S, which is the sum in column S, 4051.1, divided by the number of individuals, 25, should give the same result as that obtained by summing the means of the other five variables. The values for lines A_1 and A_2 are determined in accordance with the indicated symbols, and the values in line A_3 are obtained by subtracting the values in line A_2 from those in line A_1. At this point the value in line A_3, column S, may be used for checking the calculations. This value has been determined from the values in lines A_1 and A_2, column S, and it should equal the sum of the values for the several variables in line A_3. Thus,

$$2071.44+2195.12+210.08+2851.68+364.828=7693.148$$

It will be noted that in this table, as well as in Table 64, there are several blank spaces. For example, there are no values for lines B_1, B_2, B_3, and B_4, in column A. This is done to avoid repeating those values which appear in other columns of the table. For example, ΣAB is the same as ΣBA, and so on.

TABLE

SYSTEMATIC ARRANGEMENT OF SYMBOLS FOR DETERMINING CORRELATION

A	B	C
SUMS ΣA **MEANS** M_a	ΣB M_b	ΣC M_c
A_1 ΣA^2 A_2 $(\Sigma A)M_a$ A_3 $\Sigma A^2-(\Sigma A)M_a$ A_4 $\sqrt{\Sigma A^2-(\Sigma A)M_a}$	ΣAB $(\Sigma A)M_b$ $\Sigma AB-(\Sigma A)M_b$ $\sqrt{\Sigma A^2-(\Sigma A)M_a}$ $\sqrt{\Sigma B^2-(\Sigma B)M_b}$	ΣAC $(\Sigma A)M_c$ $\Sigma AC-(\Sigma A)M_c$ $\sqrt{\Sigma A^2-(\Sigma A)M_a}$ $\sqrt{\Sigma C^2-(\Sigma C)M_c}$
B_1 B_2 B_3 B_4	ΣB^2 $(\Sigma B)M_b$ $\Sigma B^2-(\Sigma B)M_b$ $\sqrt{\Sigma B^2-(\Sigma B)M_b}$	ΣBC $(\Sigma B)M_c$ $\Sigma BC-(\Sigma B)M_c$ $\sqrt{\Sigma B^2-(\Sigma B)M_b}$ $\sqrt{\Sigma C^2-(\Sigma C)M_c}$
C_1 C_2 C_3 C_4		ΣC^2 $(\Sigma C)M_c$ $\Sigma C^2-(\Sigma C)M_c$ $\sqrt{\Sigma C^2-(\Sigma C)M_c}$
D_1 D_2 D_3 D_4		
Y_1 Y_2 Y_3 Y_4		

To check the value in line B_1, column S, we sum the values in the preceding columns, adding ΣAB from line A_1 in column B. We include ΣAB since it is the same as ΣBA, which has been omitted to avoid repetition.

Another illustration of the use of column S for checking is given in line B_3. The various values as indicated are obtained for the several columns, including column S. The calculations may be checked by summing the values in line B_3 for columns B, C, D, and Y, adding

62

COEFFICIENTS FOR ANALYSIS OF MULTIPLE CORRELATION

D	Y	S
ΣD M_d	ΣY M_y	ΣS M_s
ΣAD $(\Sigma A)M_d$ $\Sigma AD-(\Sigma A)M_d$ $\sqrt{\Sigma A^2-(\Sigma A)M_a}\ \sqrt{\Sigma D^2-(\Sigma D)M_d}$	ΣAY $(\Sigma A)M_y$ $\Sigma AY-(\Sigma A)M_y$ $\sqrt{\Sigma A^2-(\Sigma A)M_a}\ \sqrt{\Sigma Y^2-(\Sigma Y)M_y}$	ΣAS $(\Sigma A)M_s$ Check here
ΣBD $(\Sigma B)M_d$ $\Sigma BD-(\Sigma B)M_d$ $\sqrt{\Sigma B^2-(\Sigma B)M_b}\ \sqrt{\Sigma D^2-(\Sigma D)M_d}$	ΣBY $(\Sigma B)M_y$ $\Sigma BY-(\Sigma B)M_y$ $\sqrt{\Sigma B^2-(\Sigma B)M_b}\ \sqrt{\Sigma Y^2-(\Sigma Y)M_y}$	ΣBS $(\Sigma B)M_s$ Check here
ΣCD $(\Sigma C)M_d$ $\Sigma CD-(\Sigma C)M_d$ $\sqrt{\Sigma C^2-(\Sigma C)M_c}\ \sqrt{\Sigma D^2-(\Sigma D)M_d}$	ΣCY $(\Sigma C)M_y$ $\Sigma CY-(\Sigma C)M_y$ $\sqrt{\Sigma C^2-(\Sigma C)M_c}\ \sqrt{\Sigma Y^2-(\Sigma Y)M_y}$	ΣCS $(\Sigma C)M_s$ Check here
ΣD^2 $(\Sigma D)M_d$ $\Sigma D^2-(\Sigma D)M_d$ $\sqrt{\Sigma D^2-(\Sigma D)M_d}$	ΣDY $(\Sigma D)M_y$ $\Sigma DY-(\Sigma D)M_y$ $\sqrt{\Sigma D^2-(\Sigma D)M_d}\ \sqrt{\Sigma Y^2-(\Sigma Y)M_y}$	ΣDS $(\Sigma D)M_s$ Check here
	ΣY^2 $(\Sigma Y)M_y$ $\Sigma Y^2-(\Sigma Y)M_y$ $\sqrt{\Sigma Y^2-(\Sigma Y)M_y}$	ΣYS $(\Sigma Y)M_s$ Check here

the value in line A_3, column B. This sum should equal the value in line B_3, column S. The reason for using the value in line A_3 has already been given, that is, that certain values have been omitted from the table to avoid repetition.

In order to check the values in line C_3 we sum the values in line C_3, columns C, D, and Y, adding the values in lines A_3 and B_3 in column C. The calculations may be checked at other points with the values in column S.

TABLE 63

NUMERICAL VALUES AND CORRELATION COEFFICIENTS DETERMINED FROM DATA IN TABLE 61 FOLLOWING ARRANGEMENT IN TABLE 62

	A	B	C	D	Y	S
Sums	1733	909	106	1201	102.1	4051.1
Means	69.320	36.360	4.240	48.040	4.084	162.044
A_1	122203	65207	7558	86105	7442.4	288515.4
A_2	120131.56	63011.88	7347.92	83253.32	7077.572	280822.252
A_3	2071.44	2195.12	210.08	2851.68	364.828	7693.148
A_4	45.5131	2291.8713	289.8593	3667.3955	376.8894	
		r_{ab} = .9578	r_{ac} = .7248	r_{ad} = .7776	r_{ay} = .9680	
B_1		35587	4117	46380	4098.5	155389.5
B_2		33051.24	3854.16	43668.36	3712.356	147297.996
B_3		2535.76	262.84	2711.64	386.144	8091.504
B_4		50.3563	320.7042	4057.6553	416.9955	
			r_{bc} = .8196	r_{bd} = .6683	r_{by} = .9260	
C_1			490	5190	467.0	17822
C_2			449.44	5092.24	432.904	17176.664
C_3			40.56	97.76	34.096	645.336
C_4			6.367	513.1828	52.7386	
				r_{cd} = .1905	r_{cy} = .6465	
D_1				64189	5463.5	207327.5
D_2				57696.04	4904.884	194614.844
D_3				6492.96	558.616	12712.656
D_4				80.5789	667.2658	
					r_{dy} = .8372	
Y_1					485.55	17956.95
Y_2					416.9764	16544.6924
Y_3					68.5736	1412.2576
Y_4					8.2809	

It is evident that when one has a large number of variables column S will be very useful for purposes of checking. It should be stated that for the numbers to check absolutely it is necessary that the calculation of the values in the individual columns be carried to several decimal places. The values in Table 63 have been carried to a sufficient number of decimals so that the values in column S agree with those obtained by summation. Where there are a large number of calculations it is desirable to decide the number of decimal places that will be kept. If the work is carried to only three decimal places then it must be recognized that there will be some discrepancies in checking with column S. The differences will be slight and will have no effect on the interpretations that may later be based on the calculations. If the differences as indicated by column S are considerable it indicates that there has been an error at some step in the calculations, and a rechecking will be necessary.

Having checked the calculations with the values in column S, the next step is to determine the correlation coefficients. The correlation coefficient for A with B is obtained by dividing the value in line A_3, column B, by the value in line A_4, column B. The correlation r_{ac} is obtained by dividing the value in line A_3, column C, by the value in line A_4, column C, and so on. The correlation coefficients for B with the other variables are determined by dividing the values in line B_3 by the values in line B_4, beginning with column C. The correlation coefficients for the other variables are determined in the same way, and are entered in Table 63. The standard deviations for A, B, C, D, and Y may be obtained by dividing the last number in each column by $\sqrt{N-1}$.

The next step is to determine the regression coefficients or β values from the correlation coefficients, following the directions suggested in Table 64. The steps in the calculations are made clear by the instructions given in each line of the table. Substituting the correlation coefficients from Table 63 and following the indicated steps, we have the numerical values as given in Table 65, lines 1–17. It should be noted that A correlated with A gives perfect correlation, hence the value in column A, line 1, is 1.0000. The same is true for perfect correlation in the other characters.

The operations should be followed through carefully and checked by the use of column S, as was done in Tables 62 and 63. It should

TABLE 64

SYSTEMATIC ARRANGEMENT OF SYMBOLS AND DIRECTIONS FOR DETERMINING β VALUES

DIRECTIONS	BLOCK	LINE	A	B	C	D	Y	S
Enter A correlation coefficients		1	1	r_{ab}	r_{ac}	r_{ad}	r_{ay}	$\longleftrightarrow$ Sum
Change signs	A	2	-1	$-r_{ab}$	$-r_{ac}$	$-r_{ad}$	$-r_{ay}$	
Enter B correlation coefficients		3		1	r_{bc}	r_{bd}	r_{by}	$\uparrow\longrightarrow$ Sum
Multiply Line 1 by $(-r_{ab})$, Line 2, Col. B		4		$(-r_{ab})\times r_{ab}$	$(-r_{ab})\times r_{ac}$	$(-r_{ab})\times r_{ad}$	$(-r_{ab})\times r_{ay}$	$(-r_{ab})\times$Sum
Add Lines 3 and 4		5		$({}_{bb})$	(5_c)	(5_d)	(5_y)	(5_s)
Divide Line 5 by $({}_{bb})$ and change signs	B	6		-1	$({}_{bc})$	$({}_{bd})$	$({}_{by})$	Check here
Enter C correlation coefficients		7			1	r_{cd}	r_{cy}	$\uparrow\longrightarrow$ Sum
Multiply Line 1 by $(-r_{ac})$, Line 2, Col. C		8			$(-r_{ac})\times r_{ac}$	$(-r_{ac})\times r_{ad}$	$(-r_{ac})\times r_{ay}$	$(-r_{ac})\times$Sum
Multiply Line 5 by $({}_{bc})$, Line 6, Col. C		9			$({}_{bc})\times(5_c)$	$({}_{bc})\times(5_d)$	$({}_{bc})\times(5_y)$	$({}_{bc})\times(5_s)$
Add Lines 7, 8, and 9		10			$({}_{cc})$	(10_d)	(10_y)	(10_s)
Divide Line 10 by $({}_{cc})$ and change signs	C	11			-1	$({}_{cd})$	$({}_{cy})$	Check here
Enter D correlation coefficients		12				1	r_{dy}	$\uparrow\longrightarrow$ Sum
Multiply Line 1 by $(-r_{ad})$, Line 2, Col. D		13				$(-r_{ad})\times r_{ad}$	$(-r_{ad})\times r_{ay}$	$(-r_{ad})\times$Sum
Multiply Line 5 by $({}_{bd})$, Line 6, Col. D		14				$({}_{bd})\times(5_d)$	$({}_{bd})\times(5_y)$	$({}_{bd})\times(5_s)$
Multiply Line 10 by $({}_{cd})$, Line 11, Col. D		15				$({}_{cd})\times(10_d)$	$({}_{cd})\times(10_y)$	$({}_{cd})\times(10_s)$
Add Lines 12–15		16				$({}_{dd})$	(16_y)	(16_s)
Divide Line 16 by $({}_{dd})$ and change signs	D	17				-1	$({}_{dy})$	Check here
$\beta_{yd}=-dy$		*d*				β_{yd}	$(-dy)$	
β_{yc}=Sum of two terms at its right		*c*			β_{yc}	$\beta_{yd}\times({}_{cd})$	$(-cy)$	
β_{yb}=Sum of three terms at its right		*b*		β_{yb}	$\beta_{yc}\times({}_{bc})$	$\beta_{yd}\times({}_{bd})$	$(-by)$	
β_{ya}=Sum of four terms at its right		*a*	β_{ya}	$\beta_{yb}\times(-r_{ab})$	$\beta_{yc}\times(-r_{ac})$	$\beta_{yd}\times(-r_{ad})$	(r_{ay})	

TABLE 65

NUMERICAL VALUES FROM TABLE 63, AND β VALUES OBTAINED FOLLOWING DIRECTIONS IN TABLE 64

	BLOCK	LINE	A	B	C	D	Y	S
Enter A correlation coefficients	A	1	1.0000	.9578	.7248	.7776	.9680	4.4282
Change signs		2		-.9578	-.7248	-.7776	-.9680	
Enter B correlation coefficients	B	3		1.0000	.8196	.6683	.9260	4.3717
Multiply Line 1 by (-.9578), Line 2, Col. B		4		-.9174	-.6942	-.7448	-.9272	-4.2413
Add Lines 3 and 4		5		.0826	.1254	-.0765	-.0012	.1304
Divide Line 5 by (.0826) and change signs		6		-1.0000	-1.5182	.9262	.0145	-1.5787
Enter C correlation coefficients	C	7			1.0000	.1905	.6465	3.3814
Multiply Line 1 by (-.7248), Line 2, Col. C		8			-.5253	-.5636	-.7016	-3.2096
Multiply Line 5 by (-1.5182), Line 6, Col. C		9			-.1904	.1161	.0018	-.1980
Add Lines 7, 8, and 9		10			.2843	-.2570	-.0533	-.0262
Divide Line 10 by (.2843) and change signs		11			-1.0000	.9040	.1875	.0922
Enter D correlation coefficients	D	12				1.0000	.8372	3.4736
Multiply Line 1 by (-.7776), Line 2, Col. D		13				-.6047	-.7527	-3.4434
Multiply Line 5 by (.9262), Line 6, Col. D		14				-.0709	-.0011	.1208
Multiply Line 10 by (.9040), Line 11, Col. D		15				-.2323	-.0482	-.0237
Add Lines 12–15		16				.0921	.0352	.1273
Divide Line 16 by (.0921) and change signs		17				-1.0000	-.3822	-1.3822
$\beta_{yd} = .3822$		*d*				.3822	.3822	
$\beta_{yc} = .3455 - .1875 = .1580$		*c*			.1580	.3455	-.1875	
$\beta_{yb} = -.2399 + .3540 - .0145 = .0996$		*b*		.0996	-.2399	.3540	-.0145	
$\beta_{ya} = -.0954 - .1145 - .2972 + .9680 = .4609$		*a*	.4609	-.0954	-.1145	-.2972	.9680	

be noted that no values are carried forward from column *S* in Table 63 to column *S* in Table 65. The first number in column *S* in Table 65 is the sum of the correlation coefficients given in the preceding columns in line 1. The second number appearing in column *S* for line 3 is the sum of the B correlation coefficients. Since the correlation for BA is the same as that for AB, we take this value from line 1, column B, adding the B correlation coefficients down and across. Thus we have

$$.9578+1.0000+.8196+.6683+.9260=4.3717$$

Similarly, the value in column *S* for line 7 is the sum of the C correlation coefficients and includes the correlation coefficients for AC and BC from lines 1 and 3 in column C, and the three correlation coefficients in line 7.

As already stated in regard to the checking, if the numerical values are to check absolutely it is necessary that a large number of decimals be kept in the calculations. For the calculations in this example the work has been kept to four decimal places, and at certain points in column *S* there are small differences due to the dropping of decimals. For example, when four decimals are kept in the calculations as in Table 65, the number in line 6 of column *S* does not check exactly with the result obtained by summing the values in the preceding columns in line 6. However, when six decimals are retained the results check. Where the difference is slight it is evident that no appreciable error has occurred in the preceding steps, but when a discrepancy is large a recheck of the calculations should be made.

We may now proceed to determine the β values, giving careful attention to the proper signs in all operations. In column Y, beginning with line *d*, write down in reverse order with signs changed the last values in each block, beginning with block D. In order that these steps may be readily followed the blocks A, B, C, and D are designated by writing these letters on the last line, or at the end, of the block. Following this step we have for line *d* the last value in block D, −.3822; for line *c* the last value in block C, .1875; for line *b* the last value in block B, .0145; and for line *a* the last value in block A, $-.9680$, changing the signs in each case. We now transfer to column D, line *d*, the value in column Y, line *d*, or .3822, which is β_{yd}. This number is now multiplied by the last number in each block of column D, beginning with block C and continuing in reverse order, and these products are written in lines *c*, *b*, and *a*. In obtaining these and the succeeding products the value in the last block is omitted in each case. Since this value is 1.0000 the product would be the same as the number already recorded in the first line of the reverse order.

Summing the numbers in line *c* of columns Y and D we have .1580, which is written down in line *c*, column C, and which is β_{yc}. This is multiplied by the last number in each block of column C, beginning with the last number in block B, and we have $-.2399$, which is written in line *b*, column C, and $-.1145$, which is written in line *a*, column C. Summing the numbers in line *b*, columns Y, D, and C, we have .0996, which is written in line *b*, column B, and which is the value of β_{yb}. The product of this number with the last number in block A is obtained. This is $-.0954$, and it is written in line *a*, column B. Summing the numbers in columns B, C, D, and Y we have .4609, which is the value for β_{ya}.

These same results may also be obtained from lines 1, 6, 11, and 17. Beginning first with line 17, changing the sign, we have

$$1.0000\beta_{yd}=.3822, \quad \beta_{yd}=.3822$$

From line 11 by changing the signs we have

$$1.0000\beta_{yc}-.9040\beta_{yd}=-.1875$$

Substituting the value for β_{yd} (.3822) and completing the calculations we have

$$\beta_{yc}=.1580$$

From line 6 by changing the signs we have

$$1.0000\beta_{yb}+1.5182\beta_{yc}-.9262\beta_{yd}=-.0145$$

Substituting the values for β_{yc} (.1580) and β_{yd} (.3822) and completing the calculations we have

$$\beta_{yb}=.0996$$

From line 1 we have

$$1.0000\beta_{ya}+.9578\beta_{yb}+.7248\beta_{yc}+.7776\beta_{yd}=.9680$$

Substituting the values for β_{yb} (.0996), β_{yc} (.1580), and β_{yd} (.3822) we have

$$\beta_{ya}=.4609$$

Thus, in comparing line 6 with line 1 we find that line 6 gives an equation for β values from which one value (β_{ya}) has been eliminated. Line 11 gives an equation for β values from which two values, β_{ya} and β_{yb}, have been eliminated, and line 17 gives the equation for β_{yd} after the other three β values have been eliminated.

By using these β values we may form the regression equation as was done in the previous example. This regression equation may be used for predicting yield per plant on the basis of our knowledge of the other four variables. This equation is

$$\begin{aligned}Y_{abcd}=M_y&+\beta_{ya}\frac{\sqrt{\Sigma Y^2-(\Sigma Y)M_y}}{\sqrt{\Sigma A^2-(\Sigma A)M_a}}\;(A-M_a)\\&+\beta_{yb}\frac{\sqrt{\Sigma Y^2-(\Sigma Y)M_y}}{\sqrt{\Sigma B^2-(\Sigma B)M_b}}\;(B-M_b)\\&+\beta_{yc}\frac{\sqrt{\Sigma Y^2-(\Sigma Y)M_y}}{\sqrt{\Sigma C^2-(\Sigma C)M_c}}\;(C-M_c)\\&+\beta_{yd}\frac{\sqrt{\Sigma Y^2-(\Sigma Y)M_y}}{\sqrt{\Sigma D^2-(\Sigma D)M_d}}\;(D-M_d)\end{aligned}$$

Substituting the numerical values from Tables 63 and 65 the equation is

$$\begin{aligned}Y_{abcd}=4.084&+.4609\frac{8.2809}{45.5131}\;(A-69.320)\\&+.0996\frac{8.2809}{50.3563}\;(B-36.360)\\&+.1580\frac{8.2809}{6.3687}\;(C-4.240)\\&+.3822\frac{8.2809}{80.5789}\;(D-48.040)\end{aligned}$$

Completing the necessary calculations we have

$$Y_{abcd}=.0839A+.0164B+.2054C+.0393D-5.0871$$

By substituting the observed values for A, B, C, and D for the different varieties in turn, we obtain the predicted values for Y. Since the necessary steps have already been explained in the previous problem it is unnecessary to repeat them here.

Using the correlation coefficients and β values we may now obtain the coefficient of multiple correlation from the formula

$$R^2=\beta_{ya}r_{ay}+\beta_{yb}r_{by}+\beta_{yc}r_{cy}+\beta_{yd}r_{dy}$$

Substituting in this equation the numerical values obtained we have

$$R^2=(.4609\times.9680)+(.0996\times.9260)+(.1580\times.6465)+(.3822\times.8372)=.960506$$

$$R=\sqrt{.960506}=.9801$$

With this value for R we determine to what extent yield of plant is dependent on the several variables studied, from the following formula

$$100\times1-\sqrt{1-(.9801)^2}=80.15$$

This indicates that about 80 per cent of the variability of the yield of these oat plants is dependent on the four variables under observation, leaving approximately 20 per cent unaccounted for.

Partial Correlation. We have been considering the effect on one character of several characters together. There is still another field of correlation, in which it is important to know the relation between two characters when the effect of the variation of another character or characters is eliminated, or, as it is sometimes stated, held constant.

When we determine the correlation between two variables we do so without considering the effect of other factors. We do not take any account of them. Partial correlation, on the other hand, enables us to determine the correlation between two variables after eliminating the effect of other variables that have been studied. That is, we determine the correlation between two variables independent of the variation of the other variables under observation. While the statement is often made that a certain variable or variables may be held constant, this is really impossible in most cases, but the method of partial correlation does furnish a measure of the correlation independent of the other observed variables.

The application of partial correlation may be illustrated with the material in Table 57, which has been used in the study of multiple correlation. For simplicity and to provide a general form for partial correlation we will use numeral subscripts rather than the letters as were used in multiple correlation. We will substitute for A, B, C, and Y, 1, 2, 3, and 4, and these numbers refer to the characters weight per bushel (1); weight per 100 kernels (2); weight of straw (3); and yield of grain (4).

When the correlation coefficients for simple correlation are obtained, for example r_{12}, r_{13}, r_{14}, r_{23}, and so on, we have what we may now refer to as the zero order coefficients. These are the simple correlation coefficients. The partial correlation coefficient between two variables when the effect of a third variable is eliminated is indicated by the subscript. For example, $r_{12.4}$ means that the correlation between the variables 1 and 2 is determined independent of the variation of the variable 4. Such partial correlation coefficients are referred to as first order coefficients, or partial correlation coefficients of the first order.

To obtain the partial coefficient of the first order we use the correlation coefficients of zero order, and arrange them as illustrated in the following formula

$$r_{12.4}=\frac{r_{12}-(r_{14}\times r_{24})}{\sqrt{1-r^2_{14}}\sqrt{1-r^2_{24}}}$$

It may be pointed out that the partial correlation coefficient $r_{12.4}$ is the same as $r_{21.4}$. This is evident from the following formula, remembering that r_{12} and r_{21} are identical.

$$r_{21.4}=\frac{r_{21}-(r_{24}\times r_{14})}{\sqrt{1-r^2_{24}}\sqrt{1-r^2_{14}}}$$

The arrangement of the subscripts in these formulas may serve as a guide for writing the formulas for other partial correlation coefficients. The arrangement of the subscripts of r in additional formulas is based on the subscript for the partial correlation coefficient to be determined, and should be in the following order. In the numerator we should have the first and second subscripts, the first and third subscripts, and the second and third subscripts, in that order; in the denominator we should have the first and third subscripts and the second and third subscripts. For example

$$r_{23.4}=\frac{r_{23}-(r_{24}\times r_{34})}{\sqrt{1-r^2_{24}}\sqrt{1-r^2_{34}}}$$

Substituting the necessary zero order correlation coefficients from Table 58 in the formula for $r_{12.4}$ we have

$$r_{12.4}=\frac{.629-(.654\times .604)}{\sqrt{1-(.654)^2}\sqrt{1-(.604)^2}}=.388$$

This shows the relation between 1 and 2, or between weight per bushel and weight per 100 kernels, independent of 4, the yield of grain. This simply means that if it were possible from this material to obtain a large number of varieties of oats having the same or about the same yield of grain, the correlation between weight per bushel and weight per 100 kernels for these varieties would be expected to be .388.

This meaning may be illustrated further by applying regression equations, similar to those already used, to obtain the predicted values for weight per bushel from yield of grain and the predicted values for weight per 100 kernels from yield of grain, or the predicted weight per bushel and the predicted weight per 100 kernels on the basis of yield of grain. Using the values from Table 58 in the equations

$$A_y=M_a+r_{14}\frac{\sqrt{\Sigma A^2-(\Sigma A)M_a}}{\sqrt{\Sigma Y^2-(\Sigma Y)M_y}}\quad(Y-M_y)$$

$$B_y=M_b+r_{24}\frac{\sqrt{\Sigma B^2-(\Sigma B)M_b}}{\sqrt{\Sigma Y^2-(\Sigma Y)M_y}}\quad(Y-M_y)$$

remembering that A, B, and Y are the same as 1, 2, and 4, we have

$$A_y=32.206+.654\frac{6.14}{36.54}\quad(Y-63.028)$$

$$B_y=2.569+.604\frac{1.23}{36.54}\quad(Y-63.028)$$

From these equations we have

$$A_y = 32.206 + .110Y - 6.933 = .110Y + 25.273$$
$$B_y = 2.569 + .020Y - 1.261 = .020Y - 1.308$$

Using these regression equations we obtain the predicted values for weight per bushel and weight per 100 kernels. From these predicted values we have two sets of errors of estimate, which are the differences between the observed and the calculated values for weight per bushel and the observed and calculated values for weight per 100 kernels. These errors of estimate represent the variation in weight per bushel and weight per 100 kernels after removing the effect of yield of grain.

These errors of estimate may be arranged in pairs in accordance with the variety number and the simple correlation between them determined. This correlation will be the correlation between weight per bushel and weight per 100 kernels when the effect of yield of grain is eliminated. In other words, this simple correlation coefficient obtained in this way from the errors of estimate for r_{12} should agree with the value for $r_{12.4}$ obtained from the above equation. This illustrates what is meant by partial correlation. It is the effort to obtain the relation between two characters when the effect of one or more additional characters has been eliminated.

We may proceed to make use of the above formula for partial correlation and obtain all the various first order partial correlation coefficients. It is evident that with four variables there are a number of possible first order coefficients, as follows: $r_{12.3}$, $r_{12.4}$, $r_{13.2}$, $r_{13.4}$, $r_{14.2}$, $r_{14.3}$, $r_{23.1}$, $r_{23.4}$, $r_{24.1}$, $r_{24.3}$, $r_{34.1}$, $r_{34.2}$. Since $r_{12} = r_{21}$ it is not necessary to write $r_{21.3}$, $r_{21.4}$, and so on, since they are the same as $r_{12.3}$ and $r_{12.4}$, and so on. As we are interested in obtaining the partial correlation coefficients between the different characters, first eliminating the effect of one and then of another of the characters, it is more convenient to arrange the material as suggested in Table 66, pages 206 and 207.

The first two columns of Table 66 refer to the zero order, or simple correlation coefficients. In the first column we have the subscripts, for example 12, which indicates the correlation between variables 1 and 2. In the second column is given the value that has been obtained for this correlation coefficient, .629. Using the several simple correlation values from Table 58 they are placed in column 2 of the table in accordance with the subscripts in column 1.

In the third column of the table the values for $\sqrt{1-r^2}$ are found. These values are necessary to obtain the denominator for the partial correlation coefficient, as noted in the formula above. In the fourth column we have the product term of the numerator. It will be noted in the formula that we have in the numerator the product of two correlation coefficients. This is the product of the second and third zero order correlation coefficients in each group. For example, in the first group we have the product of r_{13} and r_{23}, or $.424 \times .543$, giving .230. In the fifth column we have the value for the whole numerator, which is the product just obtained subtracted from the first correlation value in each group of the zero order correlation coefficients. Subtracting the value in the fourth column, .230, from the value for r_{12}, .629, we obtain .399. The value for the denominator in column 6 is obtained by multiplying the values for $\sqrt{1-r^2}$ given in column 3 for each group. In this case we have $.906 \times .840 = .761$. The partial correlation coefficient of the first order is now obtained by dividing the value in the fifth column by the value in the sixth column, giving in this example .524 for the first group.

The subscript for each of the first order coefficients is determined from the zero order subscripts by writing first the subscripts appearing as the first item in each group. The third subscript, which is separated from the first two by a period, is the numerical value appearing as the second number in the second and third subscripts of each group of zero order subscripts. For example, in the first group in Table 66 we have the subscripts 12, 13, and 23. It is to be noted that the third subscript in this group indicates the correlation between the second characters designated in the first two subscripts, that is, from the subscripts 12 and 13 we see that the third correlation necessary is that between 2 and 3, and therefore the third number in the first order subscript in this case is the last number of this zero order subscript. For the different partial correlation coefficients of the first order we have the subscripts and correlation coefficients as given in Table 66. It is possible to obtain other partial correlations as, for example, 31.2, 32.1, and so on,

TABLE 66

ZERO ORDER SUBSCRIPTS AND COEFFICIENTS, AND STEPS IN DETERMINING FIRST ORDER SUBSCRIPTS AND COEFFICIENTS

ZERO ORDER SUBSCRIPT 1	ZERO ORDER COEFFICIENT 2	$\sqrt{1-r^2}$ 3	PRODUCT TERM OF NUMERATOR 4	WHOLE NUMERATOR 5	DENOMINATOR 6	FIRST ORDER SUBSCRIPT 7	FIRST ORDER COEFFICIENT 8
12 13 23	.629 .424 .543	.906 .840	.230	.399	.761	12.3	.524
12 14 24	.629 .654 .604	.756 .797	.395	.234	.603	12.4	.388
13 12 32	.424 .629 .543	.777 .840	.342	.082	.653	13.2	.126
13 14 34	.424 .654 .791	.756 .612	.517	−.093	.463	13.4	−.201
14 12 42	.654 .629 .604	.777 .797	.380	.274	.619	14.2	.443

14 13 43	.654 .424 .791	.906 .612	.335	.319	.554	14.3	.576
23 21 31	.543 .629 .424	.777 .906	.267	.276	.704	23.1	.392
23 24 34	.543 .604 .791	.797 .612	.478	.065	.488	23.4	.133
24 21 41	.604 .629 .654	.777 .756	.411	.193	.587	24.1	.329
24 23 43	.604 .543 .791	.840 .612	.430	.174	.514	24.3	.339
34 31 41	.791 .424 .654	.906 .756	.277	.514	.685	34.1	.750
34 32 42	.791 .543 .604	.840 .797	.328	.463	.669	34.2	.692

but these would be the same as the values already obtained for 13.2, 23.1, and the like.

It may be desirable to continue the partial correlation study and obtain partial correlation coefficients of a higher order. Those that have just been obtained, as indicated, are known as the partial correlation coefficients of the first order. We may now obtain the partial correlation coefficients of the second order from the partial correlation coefficients of the first order. That is, for determining the partial correlation coefficients of a higher order we need the different correlation coefficients of the next lower order. Thus, for the second order partial correlation coefficients we need the coefficients of the first order partial correlation, applying them in the equation

$$r_{12.34}=\frac{r_{12.3}-(r_{14.3}\times r_{24.3})}{\sqrt{1-r^2_{14.3}}\sqrt{1-r^2_{24.3}}}$$

Substituting in this formula the values for the first order partial correlation coefficients we have

$$r_{12.34}=\frac{.524-(.576\times .339)}{\sqrt{1-(.576)^2}\sqrt{1-(.339)^2}}=\frac{.524-.195}{.817\times .941}=\frac{.329}{.769}=.428$$

This means that the relationship between 1 and 2 when the effects of 3 and 4 have been eliminated is represented by the partial correlation coefficient of .428. This may be explained in that if it were possible to have a number of varieties from this material with the same or practically the same values for the third and fourth characters, the correlation between 1 and 2 would be .428.

It may be well to point out that the formula for $r_{12.34}$ may be written in a slightly different form by a different grouping of the first order partial correlation coefficients. We have

$$r_{12.34}=\frac{r_{12.4}-(r_{13.4}\times r_{23.4})}{\sqrt{1-r^2_{13.4}}\sqrt{1-r^2_{23.4}}}=\frac{.388-(-.201\times .133)}{\sqrt{1-(-.201)^2}\sqrt{1-(.133)^2}}$$
$$=\frac{.388-(-).027}{.980\times .991}=\frac{.415}{.971}=.427$$

This affords a convenient method of checking. Other second order partial correlation coefficients may also be derived from different groupings of the first order coefficients. For example

$$r_{13.24} = \frac{r_{13.2} - (r_{14.2} \times r_{34.2})}{\sqrt{1 - r^2_{14.2}}\sqrt{1 - r^2_{34.2}}}$$

or

$$r_{13.24} = \frac{r_{13.4} - (r_{12.4} \times r_{32.4})}{\sqrt{1 - r^2_{12.4}}\sqrt{1 - r^2_{32.4}}}$$

and

$$r_{14.23} = \frac{r_{14.2} - (r_{13.2} \times r_{43.2})}{\sqrt{1 - r^2_{13.2}}\sqrt{1 - r^2_{43.2}}}$$

or

$$r_{14.23} = \frac{r_{14.3} - (r_{12.3} \times r_{42.3})}{\sqrt{1 - r^2_{12.3}}\sqrt{1 - r^2_{42.3}}}$$

For the calculations of the partial correlation coefficients of the second order it is convenient to arrange the material in a table, as was done for the first order partial correlation coefficients. For the material being studied we have arranged Table 67 on page 210.

In Table 67, the first two columns refer to the partial correlation coefficients of the first order, and in the first column the subscripts are given. The first order partial correlation coefficients from Table 66 are recorded in the second column. The other columns are determined in the same manner as was done for the first order coefficients. The subscripts and values for the second order coefficients are given in the seventh and eighth columns of Table 67. As already indicated, the interpretation of these second order partial correlation coefficients is made similarly to that for the first order partial correlation coefficients.

It is sometimes desirable to carry the calculations further and obtain the partial correlation coefficients of the third or higher orders. This may be done by extending the formula, and for the partial correlation coefficient of the third order we have the equation

$$r_{12.345} = \frac{r_{12.34} - (r_{15.34})(r_{25.34})}{\sqrt{1 - r^2_{15.34}}\sqrt{1 - r^2_{25.34}}}$$

TABLE 67

FIRST ORDER SUBSCRIPTS AND COEFFICIENTS, AND STEPS IN DETERMINING SECOND ORDER SUBSCRIPTS AND COEFFICIENTS

FIRST ORDER SUBSCRIPT 1	FIRST ORDER COEFFICIENT 2	$\sqrt{1-r^2}$ 3	PRODUCT TERM OF NUMERATOR 4	WHOLE NUMERATOR 5	DENOMINATOR 6	SECOND ORDER SUBSCRIPT 7	SECOND ORDER COEFFICIENT 8
12.3	.524		.195	.329	.769	12.34	.428
14.3	.576	.817					
24.3	.339	.941					
13.2	.126		.307	−.181	.648	13.24	−.279
14.2	.443	.897					
34.2	.692	.722					
14.2	.443		.087	.356	.716	14.23	.497
13.2	.126	.992					
43.2	.692	.722					
23.1	.392		.247	.145	.624	23.14	.232
24.1	.329	.944					
34.1	.750	.661					
24.1	.329		.294	.035	.608	24.13	.058
23.1	.392	.920					
43.1	.750	.661					
34.1	.750		.129	.621	.868	34.12	.715
32.1	.392	.920					
42.1	.329	.944					

For the partial correlation coefficient of a higher order the general equation may be written

$$r_{12.345\cdots n}=\frac{r_{12.345\cdots(n-1)}-r_{1n.345\cdots(n-1)}\,r_{2n.345\cdots(n-1)}}{\sqrt{1-r^2_{1n.345\cdots(n-1)}}\,\sqrt{1-r^2_{2n.345\cdots(n-1)}}}$$

This last equation will be understood more clearly if it is compared with the formula for the third order partial correlation coefficient. It is noted that $n-1$ appears in this general equation. In the subscript for r on the left side of the equation the letter n also appears. This indicates that the subscript may be carried to several additional numbers. The first value in the numerator is given as $r_{12.345\,\cdots\,(n-1)}$. Comparing this with the equation for the third order partial correlation coefficient it is noted that in the equation for the third order coefficient the subscript of r on the left side of the equation is 12.345 but that r in the numerator has the subscript 12.34, which is one subscript less than the subscript for r on the left side of the equation. This explains the meaning of $n-1$, since when the partial correlation coefficient is obtained as 12.345 the r values in the numerator and denominator have a subscript of one less than this. If the coefficient to be obtained is $r_{12.3456}$, then in the numerator we begin by writing $r_{12.345}$, and so on.

When it is desired to calculate one of these higher order partial correlation coefficients, tables may be arranged as before. For example, when the third order partial coefficients are to be determined, a table would be arranged using the numerical values for the second order coefficients in a manner similar to the method used with the preceding tables, and we may continue building up these higher order coefficients.

As the order of the partial correlation coefficients increases it requires an increasing number of zero order correlation coefficients. For example, for the first order partial coefficients, three zero order correlation coefficients are needed, namely, r_{12}, r_{13}, r_{23}. For the second order partial coefficients three additional zero order correlation coefficients, or a total of six, are needed. These are r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, and r_{34}. The following formula gives the number of zero order correlation coefficients needed in accordance with the order of the partial correlation coefficients.

$$\frac{(n+2)(n+1)}{2}$$

In this formula n represents the order of the partial. For example, for the third order partial we have

$$\frac{(3+2)(3+1)}{2}=10$$

This means that for the third order partial, 10 zero order correlation coefficients are needed.

Let us now consider the relation between partial correlation coefficients and regression coefficients, by determining first the regression equations for the three variables involved in the equation

$$r_{12.4}=\frac{r_{12}-(r_{14}r_{24})}{\sqrt{1-r^2_{14}}\sqrt{1-r^2_{24}}}$$

The three variables concerned here, following our notation, are 1, 2, and 4.

To obtain the β values for these regression coefficients we use the same method for obtaining β values as explained on page 182. Thus for $\beta_{12.4}$ we have

$$\beta_{12.4}=\frac{r_{12}-(r_{24}r_{14})}{1-r^2_{24}}$$

and

$$\beta_{14.2}=\frac{r_{14}-(r_{42}r_{12})}{1-r^2_{42}}$$

It is noted that in this case we are using the complete notation for β, while earlier we used an abbreviated form. These abbreviated forms have been used for convenience to save writing all of the various subscripts, which include all of the characters involved. In order that there may be a clear understanding of the complete subscripts, it may be well to write these out for one or two β values. For example, in Tables 64 and 65 we have β_{ya} and

β_{yb}. The full notations are $\beta_{ya.bcd}$ and $\beta_{yb.acd}$. The full subscripts for the other β values referred to may be written in a similar manner. In connection with the β's it may be well to point out that β_{ya} and β_{ay} are not equal numerically, since β_{ay} shows the change in A for a corresponding unit change in Y, while β_{ya} shows the change in Y for a unit change in A.

Substituting the values for the correlation coefficients indicated in the two formulas for the β values, we have

$$\beta_{12\cdot4}=\frac{.629-(.604\times.654)}{1-(.604)^2}=.368$$

$$\beta_{14\cdot2}=\frac{.654-(.604\times.629)}{1-(.604)^2}=.431$$

In a similar manner we may proceed to determine $\beta_{21.4}$ and $\beta_{24.1}$. We have the equations

$$\beta_{21\cdot4}=\frac{r_{21}-(r_{14}r_{24})}{1-r^2_{14}}$$

and

$$\beta_{24\cdot1}=\frac{r_{24}-(r_{41}r_{21})}{1-r^2_{41}}$$

Substituting the numerical values in these equations we find

$$\beta_{21.4}=.409$$

$$\beta_{24.1}=.337$$

Using the β values obtained from the first two equations we may obtain a regression equation which will give the predicted values for weight per bushel (1) when the relationship with yield of grain (4) is taken into account. Thus we have the following equation, remembering that the characters A, B, and Y are the same as 1, 2, and 4.

$$A=M_a+\beta_{12.4}\frac{\sqrt{\Sigma A^2-(\Sigma A)M_a}}{\sqrt{\Sigma B^2-(\Sigma B)M_b}}\ (B-M_b)$$

$$+\beta_{14.2}\frac{\sqrt{\Sigma A^2-(\Sigma A)M_a}}{\sqrt{\Sigma Y^2-(\Sigma Y)M_y}}\ (Y-M_y)$$

Substituting the numerical values we find

$$A=32\ 206+1.837B-4.719+.072Y-4.538=1.837B+.072Y+22.949$$

With the second set of β values we obtain a regression equation for predicting weight per 100 kernels (2) when the variation of the yield of grain (4) is taken into account. The equation is

$$B=M_b+\beta_{21.4}\frac{\sqrt{\Sigma B^2-(\Sigma B)M_b}}{\sqrt{\Sigma A^2-(\Sigma A)M_a}}\ (A-M_a)$$
$$+\beta_{24.1}\frac{\sqrt{\Sigma B^2-(\Sigma B)M_b}}{\sqrt{\Sigma Y^2-(\Sigma Y)M_y}}\ (Y-M_y)$$

Substituting the numerical values we find

$$B=2.569+.082A-2.641+.011Y-.693=.082A+.011Y-.765$$

By using $\beta_{12.4}$ and $\beta_{21.4}$ we may obtain

$$r_{12.4}=\sqrt{\beta_{12.4}\,\beta_{21.4}}=\sqrt{.368\times.409}=.388$$

This is the partial correlation coefficient and shows the relation between the correlation coefficient and the regression coefficient, and the same relationship holds whether we are dealing with simple or with partial correlation coefficients.

It is also possible to obtain the partial correlation coefficient $r_{12.4}$ from the coefficient of B (1.837) in the first regression equation above and the coefficient of A (.082) in the second regression equation, by obtaining the square root of their product. We have

$$r_{12.4}=\sqrt{1.837\times.082}=.388$$

This is the same value as obtained above.

The foregoing discussion shows how simple correlation may be extended so that by means of multiple correlation we are able to determine the effect that several variables may have on one variable. Partial correlation tends to measure the relation between two variables when the effect of one or more variables whose relationships are known have been eliminated.

It may be well to point out, however, that the methods for multiple and partial correlation have certain limitations. In the first place it is important that the number of observations studied in each instance be fairly large. A second limitation is that for these methods to give accurate results it is necessary that the regression between all variables be linear. When the regression deviates slightly from linearity the methods would still give an approximate measure for multiple and partial correlation. It is desirable to make scatter diagrams or correlation tables of the zero order correlation between the different variables to see if there is any tendency to depart from linearity. When there is a tendency to an appreciable departure from linearity, it is necessary to use methods that are adapted for such cases, and the student is referred to the treatise by Ezekiel for a full treatment of the subject.

CHAPTER IX

THE PROBABLE ERROR CONCEPT

We cannot proceed far in the study of statistical methods without realizing that all measurements and the constants derived from them are subject to a certain amount of variation, and that the reliability of a result will depend on having some estimate of the amount of variation to be expected. In dealing with the different constants as calculated in the preceding chapters, we have said nothing about this possibility. It is evident, however, that since the constants are determined from data obtained by measurement or counting, such data are subject to a certain amount of variation. There may be variation in such a simple process as taking a measurement on the height of a wheat plant. It is likely that the same plant may be measured by two investigators and the results may be slightly different due to the fact that the measure was not read in exactly the same way. It is impossible to attain absolute accuracy in measurements or weights, so that the results obtained from two or more lots of the same material may not give the same mean or average.

It is also evident that the reliability of a result is related to the number of observations on which the result is based. We often hear the expression that two guesses are better than one, and in statistical analysis this may be extended to mean that a constant calculated from many observations of similar kind is more reliable in general than one calculated from only a few observations.

Errors of Observation. Let us illustrate the variation that may occur in results by considering a chemical problem. Suppose that a number of persons were to determine the amount of silver in a solution of silver nitrate. Even if they were well-trained chemists and used the most approved methods with accurate balances, the final determinations as to the amount of silver present in the solution will differ. Such differences will be due in part to the samples that each experimenter may take, and we may say that sampling causes part of the variation. Variation will also be due to the technique followed by the different experimenters. Even if several determinations were made by one investigator, his results also would show a certain amount of variation. The average of the several determinations, however, may give a fair measure of the amount of silver in the solution, but it is important that this average be based on sufficient determinations.

What is true of this chemical experiment is true also of other types of chemical experiments, physical measurements, measurements of plant characters, results obtained from surveys, data from social or economic sources, and the like. If a large number of measurements have been made on similar material many of them may agree very closely but others may show a considerable amount of variation. The differences that are caused by any uncontrollable factor are called deviations, and may be termed errors of observation. They occur and affect most results, and they cannot be entirely eliminated, but their effect may be minimized by the use of large numbers of observations. It is possible to use the most refined methods in experimental work and yet many uncontrollable factors intervene to cause results to deviate more or less from each other and from the mean of all of the results. Such variation is expected, and in fact one often finds that when experimental results agree too closely they arouse a certain feeling of distrust.

These errors of observation must not be confused with mistakes in measuring and recording, or in the calculations, nor should the fact that we expect variations in a certain result be an excuse for inaccuracies of measurement. In all operations one should be as exact as possible, but even so there will still be many uncontrollable factors that will affect the results and produce errors of observation.

Errors of observation may be illustrated by the example of a marksman shooting at a target. The object is to hit the center of the target, but there will be a certain amount of deviation from the center depending on the accuracy of the marksman. The experimenter, like the marksman, is attempting to measure or obtain a true value and, like the marksman, each determination he makes may miss the true value more or less, depending on the variability of the material and the methods of measuring and recording the results.

Let us consider as an experiment a contest between two men shooting at a target. Each marksman has ten shots or chances. The deviations of the shots of the first marksman, A, from the center of the target are indicated in column A of Table 68, and the deviations of the shots of the second marksman, B, are given in column B of the table.

TABLE 68

ASSUMED DEVIATIONS FROM THE CENTER OF A TARGET IN A CONTEST BETWEEN TWO MARKSMEN

DEVIATIONS A	DEVIATIONS B
0	2
3	3
4	2
2	4
3	3
4	2
5	3
0	4
5	2
1	2
27	27

These failures to hit the center of the target may be considered as errors of observation. Marksman A hits the center of the target twice, as indicated by the two deviations of zero in column A, but he also misses the center of the target twice by five units each time. Marksman B does not hit the center in any of his shots, yet he does not miss the center by as many units as does marksman A.

We may now compare these results. Summing the deviations of A and B we find that the sum in each case is 27. Since there are ten experiments or ten shots we divide 27 by 10, giving an average of 2.7 for each of the marksmen. This result alone would indicate that the marksmen are equal in their ability to shoot, or if we consider it as covering experiments in general we might say that the results for A and B are the same.

Let us consider, however, the size of the target that is necessary to include all the shots for A. It is seen that in two of his attempts he deviates from the center by five units. This means that he must have a larger target to include all of his shots than does B, whose greatest deviation from the center of the target is four units. Instead of using the average of the deviations a better measure of the accuracy of A and B may be obtained by the root-mean-square as measured from the center of the target. We find that the sum of the squares of A's shots is 105 and the sum of the squares of B's shots is 79. Taking the mean, we have 10.5 for A and 7.9 for B, and extracting the square root we have 3.24 for A and 2.81 for B. We would conclude that B is a more accurate marksman than A since his shots show less variability. Similar results may arise in experimental work, and the reliability of the conclusion drawn will be dependent on the variability of the results.

Let us consider another simple experiment which will illustrate the meaning of errors of observation, or the tendency of the individuals to deviate from some central value. For this experiment we utilize the popular practice of tossing coins. If we consider first the tossing of one coin, one side of which may be designated as heads and the other as tails, we know that when this coin is tossed on a table there are only two possibilities for it to come to rest, that is with either the head side up or the tail side up. Now what is true of one coin is true of several coins, and when several are tossed on a table each coin has the possibility of coming to rest with either side up.

We may suppose that eight coins are shaken thoroughly in the hand or in a cup and tossed on a table. While we may expect half of them to come to rest with heads up and half with tails up, this does not hold true all of the time, and there is a certain amount of deviation. It is entirely possible if one continues the tossing that occasionally all the coins may come to rest with heads up, or all may come to rest with tails up. It is also possible that only one of the eight will come to rest with heads up, and again we may find two heads out of the eight, or three heads, and so on. Thus there are several possible variations ranging from no heads showing to all eight heads showing.

If we continue to toss eight coins many times we may expect to find results similar to those shown in Table 69, page 220, in which are recorded the number of heads obtained by tossing eight coins 2000 times.

While, as expected, four heads and four tails occur most frequently, we also find that three heads and five heads occur very often, and that two and six heads, one and seven heads, and zero and eight heads occur with less frequency. Such an experiment gives results that tend to approach a symmetrical distribution. If the results came out exactly as expected we should have the number of heads as indicated in the last column of Table 69. With such an experiment we would expect the values given by the expansion of the binomial 2000 $(\frac{1}{2} + \frac{1}{2})^8$, but it is seen that the observed numbers in the second column show some variation from those in the last column. That is, even with such a simple experiment as the tossing of coins, we find considerable deviation from expectancy. Even greater deviations may be expected in results obtained from biological, social, or economic statistics.

TABLE 69

DISTRIBUTION OF NUMBER OF HEADS
OBTAINED BY TOSSING EIGHT COINS
2000 TIMES

NUMBER OF HEADS POSSIBLE	NUMBER OF TIMES OBTAINED	NUMBERS OBTAINED BY EXPANDING THE BINOMIAL 2000 $(\frac{1}{2}+\frac{1}{2})^8$
0	11	7.8125
1	62	62.5000
2	196	218.7500
3	421	437.5000
4	574	546.8750
5	467	437.5000
6	203	218.7500
7	55	62.5000
8	11	7.8125

The results given in Table 69 show that there is a tendency for the errors of observation to be symmetric, with the greatest number of observations at the center, or at class 4. If the tossing were continued so that we would have a very large number in the population, then the observed values would approach more and more closely to a perfect symmetrical distribution, similar to the distribution in the last column of the table. This is true with errors of observation in general, that is, most errors of observation tend to form a symmetrical curve, especially if a large number of such errors are available. With a small number in the population the distribution may not be symmetrical. The symmetrical distribution resulting from errors of observation is referred to as a normal distribution, and such a normal distribution is often referred to as a normal curve of error. For the analysis of the results of errors of observation we may use the methods of analysis that have been developed for the normal curve of error.

Since there is this tendency for measurements and observations to vary, it is evident that the results obtained from any biometrical analysis are subject to a certain amount of variation due to errors of sampling. That is, in studying such a character as the height of men in a certain city or of a particular race, it is not possible for us to measure all of the individuals and therefore the number with which we deal is merely a small sample drawn from the whole population. If one measures the height of 2000 men and finds the average to be 67.8 inches, this does not mean that the height of all of the men in the city, or all of the men of a particular race, would average exactly 67.8 inches. That is, while the first sample of 2000 men may show an average height of 67.8 inches, it is entirely possible that another sample of 2000 will give a mean value slightly different from the first, and the same will be true of additional samples. Therefore we must not treat the result obtained from the sample of 2000 men as a definite value for all the men. It is merely an average value that is definite for this study and another sample may give a slightly different result, or the addition of even a few more men to the sample of 2000 may change the result.

We cannot know definitely the average height of all the men in a city or of a certain race unless all of the individuals are measured, and it is usually impossible to do this. It may be well to point out that in the usual statistical analysis we are endeavoring to obtain results which represent the true value that would be found if it were possible to make observations on all of the individuals or cases. In other words, we may not know or ever obtain exactly the true value, but by averaging the results of many observations or experiments we may obtain the most probable value which approaches the true value.

Let us illustrate the differences that may occur due to the deviations of the individual observations or due to the fact that we are able to measure or observe only a portion of the total individuals. We will refer again to the experiment of tossing coins. Suppose eight coins are tossed 20 times and the number of heads recorded, giving the following distribution:

NUMBER OF HEADS POSSIBLE	NUMBER OF TIMES OBTAINED
0	0
1	1
2	2
3	4
4	4
5	7
6	2
7	0
8	0

The mean of this distribution is 4.00. Now, suppose we make one more toss and add it to these 20. The number of heads turned up is 3, and the mean of this distribution of 21 tosses is found to be 3 95. It is seen that the addition of one more observation has had the effect of changing the mean from 4.00 to 3.95. This shows that the mean may be changed by the addition of one or more individuals, depending on the deviation that the additional observations may show from the mean. If the additional toss had been the same as the mean of the first 20, the mean would have remained at 4.00, but as it differed by one unit from 4.00 its effect was to reduce the value of the mean.

To show the change in the value of the mean that occurs in several samples drawn from the same kind of material, we may consider again the experiment of tossing coins. Suppose that we conduct four more experiments in which we toss eight coins 20 times. The results of these five experiments are given in Table 70, page 223.

From the first experiment we obtained a mean value of 4.00. If there were no errors due to sampling we would expect that the means of the other four experiments would also be 4.00, but we find that they differ by various amounts. The means of the other four experiments are 4.20, 3.95, 4.10, and 3.95. Thus we see that with an experiment as simple as the tossing of coins there is considerable variation so far as the mean value is concerned.

TABLE 70

RESULTS OF TOSSING EIGHT COINS 20 TIMES IN FIVE EXPERIMENTS

NUMBER OF HEADS POSSIBLE	NUMBER OF TIMES OBTAINED IN EACH EXPERIMENT				
	FIRST	SECOND	THIRD	FOURTH	FIFTH
0	0	0	0	0	0
1	1	0	1	1	0
2	2	3	1	2	4
3	4	4	6	3	4
4	4	5	3	7	3
5	7	3	8	3	7
6	2	4	1	3	2
7	0	1	0	1	0
8	0	0	0	0	0
MEAN	4.00	4.20	3.95	4.10	3.95

Standard Error and Probable Error. If we find such variation in this experiment, it is evident that there will be a considerable amount of variation in the means and other constants when samples are drawn from biological, social, or economic statistics. With such a simple experiment as the tossing of coins we may repeat the operation a large number of times and finally obtain a result that approaches very closely the true result, but with the usual type of experiment it is impossible to have a large number of replications. Therefore it is necessary to have some means of indicating the amount of variation that may be expected in a constant or, in other words, to be able to indicate the reliability of a constant. In order to satisfy this condition it is necessary to add to a constant, such as the mean, standard deviation, coefficient of correlation, and the like, a value that will indicate this variability or measure the reliability of these constants. We have such a measure in either the standard error or the probable error. At the present time in certain types of investigation, especially in the analysis of field experiments and similar studies, the standard error is being used rather generally, but the probable error still has a place in statistical analysis. Standard errors may be converted to probable errors by multiplying the standard error by the constant .6745.

The use of the standard error or the probable error to indicate the expected variability or to measure the reliability of an experiment may be illustrated with the first experiment of the tossing of 20 coins. in Table 70 The mean of this experiment is 4.00 and the standard deviation is 1.34. This gives us some idea of the amount of variation we may expect from single tosses, since the standard deviation is the standard error of a single toss in this experiment and may be used to denote the variability that may be expected to occur in the mean if we made other experiments with similar material. As already stated, the probable error is obtained by multiplying the standard error by the constant .6745, and for this experiment we have 1.34 × .6745, giving a value for the probable error of .90. This is referred to as the probable error of a single determination, which in this case is a single toss. and is represented by the formula

$$P.E._s = \pm .6745\sigma$$

Now, as we found by experience, we do not know whether a single toss will give us a value above or below the mean, and since either is likely to occur the practice is to precede the constant .6745 by the plus and minus sign (±) and the probable errors as determined are preceded by this ± sign, indicating that it is equally possible for the error to be above or below the central value, or the mean.

The standard error and probable error give some idea of the possible deviations from the mean that may be expected in future experiments. In the present experiment either of these constants would give some idea of the deviations that may be expected if we continued the tossing of coins. As already stated $\pm .6745\sigma$ is the probable error of a single observation, or toss in this case. It is possible to obtain the probable error of the mean of this experiment by making use of the number of determinations. It has been stated that the reliability of a result depends on the variability that the several items show and on the number that have been observed. Therefore, by considering the variability and the number of observations, we may obtain the probable error of the mean from the formula

$$P.\,E._M = \pm .6745 \frac{\sigma}{\sqrt{N}}$$

This shows that while the reliability depends on the number of observations it does not increase directly with the number of individuals observed but increases with the square root of the number. Substituting in this equation the standard deviation and N for the first experiment in Table 70 we have the probable error of the mean, or

$$P.\,E._M = \pm .6745 \frac{1.34}{\sqrt{20}} = \pm .20$$

This gives an indication of the variation we may expect in the mean if we continue to make other experiments of a similar kind.

We are able now to make some prediction as to the limits within which or without which we may expect the mean to fall. Since the standard deviation when multiplied by the constant ± .6745 gives us the quartile deviation for symmetrical distributions, and since we have already learned that the quartile deviation is such that one-half of the observations are included within the limits of the quartile deviation we may now apply this idea to the probable error of a mean.

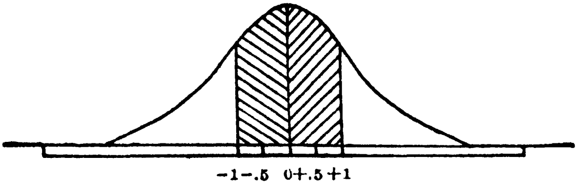


FIG. 28. Illustrating area of curve included when a value equal to 1 × $P.E.$, is measured off on each side of the mean.

The quartile deviation, as already stated, is equal to the probable error of a single determination, and in Figure 28, page 225, is shown the area of the curve that is included when the value of $P.E._s$, or the quartile deviation, is measured off on each side of the mean. Thus we see that when the value of $1 \times P.E._s$ is measured from the mean in both directions we include one-half of the area of the curve, or applied to a frequency distribution we include one-half of the observations. The same is true if the observations are means, or errors of observation.

We may therefore expect on the basis of our first experiment that, repeating the experiment a large number of times, one-half of the means will fall within the limits indicated by the probable error, or the amount as calculated from the above formula. On the basis of this relation we expect that one-half of our means will fall within the limits of 4.00 ± .20, or from 3.80 to 4.20, inclusive. This is on the assumption that we have a large number of experiments or observations, but may not hold exactly for small numbers.

In order to compare observation with theory, the experiment of tossing eight coins 20 times was continued until 194 such experiments were obtained. The mean of each experiment was calculated and these means are recorded in Table 71, page 227.

From the first experiment we predicted that the probable error of a mean of a similar experiment should be ± .20. This prediction was based on 20 tosses, and we find that when the means of 194 similar experiments are obtained and the probable error of a single determination (in this case a single mean) is calculated from $\pm .6745\sigma$ the probable error is ± .20, the same value as predicted from the first experiment.

Now, in accordance with our prediction we expect one-half of the means in Table 71 to fall within the limits of 4.00 ± .20. If we count the means that lie within these limits we find 97, and there are also 97 outside these limits. As a rule such close agreement is not expected, but in this case there is an exact agreement between theory and observation—that the quartile deviation or the probable error of a single determination when measured from the mean includes one-half of the observations. We see from the result obtained how it is possible on the basis of one experiment to determine the limits within which or without which we may expect our result to fall if we were able to continue to make a large number of similar determinations.

TABLE 71

DISTRIBUTION OF MEAN NUMBER OF HEADS OBTAINED FROM 194 EXPERIMENTS OF TOSSING EIGHT COINS 20 TIMES

3.40	3.70	3.85	4.00	4.15	4.30
3.40	3.70	3.85	4.00	4.15	4.30
3.45	3.70	3.85	4.00	4.20	4.35
3.45	3.70	3.85	4.00	4.20	4.35
3.45	3.70	3.85	4.00	4.20	4.35
3.50	3.70	3.85	4.00	4.20	4.35
3.50	3.75	3.85	4.00	4.20	4.35
3.50	3.75	3.85	4.00	4.20	4.35
3.55	3.75	3.90	4.00	4.20	4.35
3.55	3.75	3.90	4.05	4.20	4.40
3.55	3.75	3.90	4.05	4.20	4.40
3.55	3.75	3.90	4.05	4.20	4.40
3.55	3.75	3.90	4.05	4.20	4.40
3.60	3.75	3.95	4.05	4.20	4.40
3.60	3.75	3.95	4.05	4.20	4.40
3.60	3.75	3.95	4.05	4.20	4.40
3.60	3.75	3.95	4.05	4.20	4.40
3.60	3.75	3.95	4.05	4.20	4.45
3.65	3.75	3.95	4.05	4.20	4.45
3.65	3.75	3.95	4.05	4.20	4.45
3.65	3.80	3.95	4.05	4.20	4.45
3.65	3.80	3.95	4.05	4.25	4.45
3.65	3.80	3.95	4.10	4.25	4.45
3.65	3.80	3.95	4.10	4.25	4.50
3.65	3.80	3.95	4.10	4.25	4.50
3.65	3.80	3.95	4.10	4.25	4.50
3.65	3.80	4.00	4.10	4.25	4.50
3.65	3.85	4.00	4.15	4.25	4.50
3.65	3.85	4.00	4.15	4.30	4.55
3.65	3.85	4.00	4.15	4.30	4.55
3.65	3.85	4.00	4.15	4.30	4.60
3.65	3.85	4.00	4.15	4.30	4.60
3.65					4.60

$M = 4.00$
$S.D. = .29$
$P.E._1 = \pm 20$

The prediction that may be made is dependent on the fact that the first experiment is based on sufficient observations, and on a sample that is fairly representative of the condition we are investigating. If the first experiment is based on inadequate numbers or on material that does not represent the whole population, then the prediction will not be so exact. If the first experiment in the tossing of coins had been one that was very abnormal, we would not obtain the close relation between theory and prediction that is found here, but the illustration has been used to show the relation between one experiment and many of a similar kind.

The reliability of constants, such as the mean and standard deviation, is therefore indicated by the probable error, and it must be evident that the smaller the error the more reliable is the result. If we have the values 40±4 and 40±1 for two means, we think of the latter as being more reliable than the former. This is true since we would expect the true mean in the case of the latter experiment to fall within the limits of 40±1 one-half of the time, while with the former the true mean would fall within the limits of 40±4 one-half of the time. We would therefore place more confidence on the mean of 40±1.

In practice it is preferable to write the constant with its probable error rather than to give merely the constant. That is, for the experiment of the tossing of coins the mean is better expressed as 4.00±.20, rather than 4.00.

Let us consider another type of experiment. A large number of plots of equal size were sown to the same variety of soy beans, and the yields were obtained from each plot. Samples containing ten plots each were made up at random, and the means of these samples are recorded in Table 72.

TABLE 72

DISTRIBUTION OF MEANS OBTAINED FROM SAMPLES FROM THE SAME VARIETY OF SOY BEANS

179.8	212.0	221.9	234.1
181.5	212.1	222.8	236.5
196.1	212.2	222.9	236.7
196.4	213.7	223.2	237.1
201.7	214.3	223.9	237.7
201.7	216.3	224.1	237.9
202.7	216.5	224.3	238.1
205.9	216.5	224.5	239.8
207.2	217.3	224.7	240.5
207.4	217.6	225.0	241.6
208.3	217.8	225.9	242.5
208.8	218.7	229.8	244.9
209.8	218.9	229.8	247.3
209.8	219.1	230.5	249.2
210.6	220.1	232.0	249.2
211.4	220.9	232.2	251.9
211.5	221.4	233.0	253.3
211.7	221.4	233.9	262.9

The mean of these means is found to be 222.7 and the standard deviation is 15.92. Obtaining the probable error for a single determination we have

$$P.E._s = \pm .6745 \times 15.92 = \pm 10.74$$

Table 73

Yields of 240 Plots All Sown to the Same Variety of Oats

35.6	57.6	61.2	65.0	69.2	73.2
42.2	58.0	61.2	65.4	69.4	73.4
44.6	58.4	61.6	65.6	69.4	73.4
46.8	58.4	61.6	65.6	69.4	73.6
49.0	58.4	61.6	65.6	69.6	73.6
49.4	58.4	61.6	66.0	69.6	73 6
49.4	58.4	62.0	66.0	69.6	73.8
49.6	58.4	62.0	66.0	69.8	74.0
49.6	58.4	62.0	66.2	69.8	74.4
49.8	58.6	62.2	66.4	69.8	74.4
50.0	58.6	62.4	66.4	69.8	75.2
51.0	58.6	62.4	66.6	70.0	75.2
51.4	58.6	62.6	66.6	70.0	76.0
51.6	58.8	62.6	66.6	70.0	76.0
51.8	59.0	62.6	66.8	70.0	76.2
52.2	59.0	62.6	67.0	70.4	76.4
52.4	59.0	62.8	67.0	70.4	76.6
52.6	59.0	62.8	67.0	70.4	76.8
52.6	59.2	63.0	67.4	70.8	76.8
52.8	59.4	63.0	67.4	70.8	77.2
53.0	59.4	63.2	67.4	71.0	77.2
53.0	59.6	63.4	67.4	71.0	78.0
53.4	59.6	63.6	67.6	71.0	78.0
54.2	60.0	63.6	67.6	71.0	78.6
54.2	60.0	63.8	67.6	71.0	79.0
55.2	60.0	64.0	67.8	71.2	79.4
55.2	60.0	64.0	68.0	71.2	79.6
55.2	60.4	64.0	68.0	71.2	79.6
55.4	60.4	64.2	68.0	71.4	79.8
55.4	60.4	64.2	68.0	71.8	80.0
55.6	60.4	64 4	68.2	72.0	80.4
56.2	60.4	64.4	68.4	72.0	80.4
56.6	60.4	64.4	68.4	72.0	80.6
56.8	60.4	64.6	68.4	72.2	80.8
56.8	60.8	64.6	68.4	72.2	81.4
57.0	60.8	64.6	68.6	72.4	82.0
57.0	61.0	64.8	68.8	72.6	82.2
57.0	61.0	64.8	69.0	72.8	82.6
57.2	61.2	64.8	69.0	73.0	83.2
57.6	61.2	65.0	69.2	73.2	83.4

$P.E._s$ obtained by counting $= \pm 5.75$

$P.E._s$ calculated from $.6745\sigma = \pm 5.68$

With this probable error of a single determination, which is the probable error of a single mean, we may mark off the limits on each side of the mean indicated by the plus and minus signs, and we have

$$222.7-10\ 74=211.96$$
$$222.7+10.74=233.44$$

If we count the means within these limits we find that there are 35 within the limits, and 37 without. From theory we expect 36 to 36, so again the observation agrees closely with theory and this illustrates the meaning of the probable error of a single determination.

Probable Error by Counting. We have already learned that the quartile deviation is equal to $P.E._s$, and since this is so it should be possible to determine the probable error of a single determination by counting. If all of the individual measurements are arranged in order according to size, beginning with either the largest or smallest item, then by counting we may determine the individual that divides the population so that one-fourth of the individuals fall below this value and three-fourths are above this value. This may be illustrated with the data in Table 73, page 229, which are the yields of 240 plots of oats all sown to the same variety, arranged in ascending order of yield.

As there are 240 individuals in the population, the first quartile must be so taken that 60 individuals fall below it Therefore the first quartile is between the 60th and 61st individual, and the average of these two individuals is taken. We have

$$\frac{59.4+59.4}{2}=59.4$$

This is the value of Q_1. Similarly the value of Q_3, the point below which three-fourths of the individuals lie, is

$$\frac{70.8+71.0}{2}=70.9$$

Then from the formula for the quartile deviation

$$Q=\frac{Q_3-Q_1}{2}$$

we have

$$Q=\frac{70.9-59.4}{2}=5.75$$

Since Q is equal to $P.E._s$, therefore the value of $P.E._s$ determined by counting is ±5.75. Now if the mean and standard deviation are determined in the usual way we have

$$M = 65.1$$
$$\sigma = 8.42$$

and $P.E._s$ is equal to

$$\pm .6745 \times 8.42 = \pm 5.68$$

The value of $P.E._s$ obtained by counting, ±5.75, agrees very well with that obtained from the calculations, ±5.68.

The mean is 65.1, and when $P.E._s$, ±5.68, is subtracted from and added to the mean we have the limits 59.42 and 70 78. Counting the individuals within these limits we find 117, and there are 123 without these limits. This agrees very well with 120 to 120, which would be expected from theory. Slight deviations from the theoretical expectation may occur, but they are not large enough to discredit conclusions that may be drawn. When the population is large and the observations agree closely with the symmetrical or normal type of distribution, wide deviations will probably be less frequent than with small populations and asymmetrical distributions.

Now if twice the value of $P.E._s$ is taken ($5.68 \times 2 = 11.36$) and new limits obtained by subtracting and adding this value from and to the mean, 65.1, the limits are 53.74 and 76.46. The number of individuals within these limits is 193 and there are 47 individuals outside these limits. This is a ratio of 193/47, or 4.11 to 1. We may now proceed to determine the number that may be expected if the population were large and the distribution were strictly symmetrical, and this leads to the consideration of the calculation of odds.

Calculation of Odds. In order to express results in terms of probability, or expectancy, we calculate odds on the assumption that the observations would give a symmetrical curve provided there were sufficient numbers. Appropriate tables of probability values are available to aid in the calculation of odds, and an abbreviated form of such a table appears as Table VI in the Appendix. The values in this table are based on an assumed area of 100000, and from these values it is possible to determine what part of the area of the curve is included for any distance from the mean, by obtaining first the ratio of the distance, x, to its standard deviation, or $\frac{x}{\sigma}$.

The values in the first column, $\frac{x}{\sigma}$, are the ratios that may be obtained by dividing any distance, x, by the standard deviation that has been determined for the data under observation. For example, if the standard deviation of a distribution is 1.5 and x is taken as .9, we have the ratio .9/1.5, or .6. For this ratio of .6, in the first column of the table, we find that the corresponding area is 22575, which is the area of the curve that is included between the mean and the value of x (.9) on one side of the mean. On both sides of the mean the area is twice this value, or $2 \times 22575 = 45150$, so when $\frac{x}{\sigma} = .6$ the area included within these limits is about 45 per cent. From the ratio of $\frac{x}{\sigma}$ it is possible to determine from the table what part of the total area of the curve would be included for any particular ratio.

As another illustration we may take the $\frac{x}{\sigma}$ ratio of .6745. If we look in column $\frac{x}{\sigma}$ for the value .67 and then read on this line in the column with heading 4, we find the value 24984. This indicates that for the ratio of $\frac{x}{\sigma}$ of .674 the area of the curve that would be included equals 24984. Now the value we are seeking is the ratio of $\frac{x}{\sigma}$ of .6745, and it is therefore necessary to interpolate. For $\frac{x}{\sigma}$ of .675 we find on the same line in the column with heading 5 that the area is 25016. Interpolation may be done as follows:

$$\frac{x}{\sigma} = .675 \quad \text{area} = 25016$$

$$\frac{x}{\sigma} = .674 \quad \text{area} = 24984$$

The differences are

$$.001 \qquad 32$$

The value we wish is .6745, or one-half the distance between .674 and .675, so we take one-half the difference between the two areas, or one half of 32, giving 16. This is added to the lower value, 24984, giving 25000. Thus we see that when $\frac{x}{\sigma}$ equals .6745 one-fourth the area of the curve is included and since the error of observation is as likely to be above as it is below the mean we measure off this value on both sides of the mean, giving a total area of 50000. Since this is equal to one-half the area of the curve, we see why we expect one-half of the individuals in a series of observations to fall within the limits of $1 \times P.E._s$, and one-half to fall without, leading to odds of 1 to 1. This is illustrated graphically in Figure 28.

Suppose for a distribution the value of x is such that when it is divided by the standard deviation of the distribution, or for $\frac{x}{\sigma}$, the ratio is 1. Then from the table the area of the curve that is included is 68268. This means that for any distribution where $\frac{x}{\sigma} = 1$ and a distance equal to the standard deviation is marked off on each side of the mean about 68 per cent of the individuals are included.

We have used $\frac{x}{\sigma}$ in determining odds, and the question may arise as to how the table of probability values giving the area of the curve for $\frac{x}{\sigma}$ may be used to calculate odds when we are dealing with the probable error. A simple illustration may help to clear up this point and show that either the standard deviation or the probable error may be used with these tables.

Suppose that an x value of 3, as measured from the mean, has been obtained, and the standard deviation is 1.5. We may call this x value the difference, or deviation. The value of $\frac{x}{\sigma}$ in this case is 2.0, and we would therefore seek out the area corresponding to an $\frac{x}{\sigma}$ value of 2.0.

If we choose to use the probable error the standard deviation will have been multiplied by the constant .6745, giving 1.012. For the value of $\frac{x}{P.E.}$ we would then have 3/1.012, or 2.964. Thus it is seen that when the standard deviation is used a ratio value of 2.0 is obtained, but when the probable error is used the ratio value is 2.964. Since the probability values in the table are determined on the basis of $\frac{x}{\sigma}$, when we use the probable error and obtain a ratio of $\frac{x}{P.E.}$ it is necessary to multiply this ratio by .6745 to give the same ratio that would have been obtained had the standard deviation been used. The ratio 2.964 multiplied by .6745 gives 1.999, and if more decimals were retained this would equal 2.0, the same value obtained when using the standard deviation. Therefore, when obtaining the ratio of any difference, or any value of x, to its probable error, it is necessary to multiply the ratio by the constant 6745 before determining the odds from the table of probability values.

We may now determine from the values in the probability table the odds when the $\frac{D}{P.E.}$ ratio is 2.0, or twice the value of $P.E._{s}$. Following the discussion above we have $.6745 \times 2 = 1.3490$, and since the table is given in an abbreviated form it is necessary to interpolate for this value. From the ratios of $\frac{x}{\sigma}$ nearest to 1.3490 we have

$$\frac{x}{\sigma} = 1.350 \quad \text{area} = 41149$$

$$\frac{x}{\sigma} = 1.329 \quad \text{area} = 40808$$

The differences are

$$.021 \qquad 341$$

The value for which we wish to determine the area is 1.3490, and the difference between this and 1.329 is .020. Now by direct interpolation we have

$$\frac{.020}{.021} \times 341 = 325$$

This is added to the lower area value, 40808, and we have 41133. Again on the assumption that the error of observation may occur above or below the mean, we double the area, giving 82266. This is subtracted from the total area, or 100000–82266, leaving 17734. This shows the proportion of the total area that would be included within twice the value of $P.E._{s}$, and the total excluded. This is illustrated diagrammatically in Figure 29, page 235. The odds are obtained from the ratio 82266/17734, or 4.64 to 1. This ratio is to be compared with that obtained from the oat problem, where the ratio was found to be 4.11 to 1. The expected value is higher than that obtained from observation.

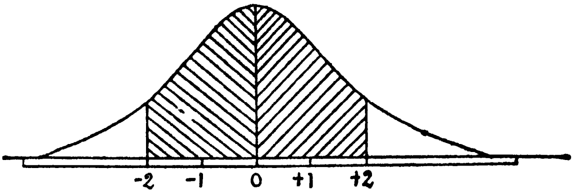


FIG. 29. Illustrating area of curve included when a value equal to $2 \times P.E._s$ is measured off on each side of the mean.

We may determine the odds expected when $\frac{D}{P.E.}$ is 3.0, or $3 \times P.E._s$, following the steps just outlined. We first obtain $3 \times .6745$, or 2.0235. It is necessary again to interpolate, and we have from the ratios of $\frac{x}{\sigma}$ nearest 2.0235

$$\frac{x}{\sigma} = 2.040 \quad \text{area} = 47932$$

$$\frac{x}{\sigma} = 2.019 \quad \text{area} = 47826$$

The differences are

$$.021 \qquad 106$$

The value for which we are interpolating is 2.0235, and subtracting 2.019 from this we have .0045. Then

$$\frac{.0045}{.021} \times 106 = 23$$

This is added to the lower area value, 47826, giving 47849. Doubling this value, for the same reason as before, we have 95698 as the portion of the curve that we would expect to be included within the limits of $3 \times P.E._s$. The difference between this and 100000, or 100000–95698 = 4302, is the amount left outside of these limits. This is illustrated diagrammatically in Figure 30. Obtaining the ratio 95698/4302 we have 22.25, or the odds are 22.25 to 1.

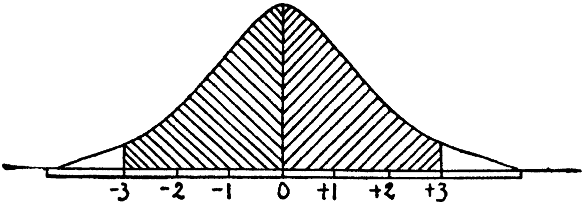


FIG. 30. Illustrating area of curve included when a value equal to $3 \times P.E._s$ is measured off on each side of the mean.

In this case we would interpret the odds as being 22.25 to 1 against a deviation being found as great or greater than three times the probable error of a single determination due to errors of sampling or chance variation alone. In other words, we would expect that once out of 23.25 times (22.25 + 1) a deviation as great as three times the probable error of a single determination will occur due to random variation, or due to chance.

We may continue the calculations for four times the probable error, and by interpolating in the same manner as before we find that the odds are 142.27 to 1 against a deviation of four times the probable error of a single determination occurring due to random variation.

The methods given show how odds may be calculated. A short table giving odds for the different ratios of $\frac{D}{P.E.}$ is given as Table 74, and a more extended table appears as Table VII in the Appendix. For the calculation of odds a more complete table is given in

TABLE 74

Table of Odds for $\frac{D}{P.E.}$ Values, Calculated from the Probability Values in Table VI in the Appendix

Difference from the Mean in Terms of the Probable Error, or $\frac{\text{Deviation}}{\text{Probable Error}}$	Difference Between Two Results in Terms of the Probable Error of Each Result*	Odds Against Such a Difference Occurring Due to Chance
1.00	1.41	1.0 : 1
1.25	1.77	1.5 : 1
1.50	2.12	2.2 : 1
1.75	2.47	3.2 : 1
2.00	2.83	4.6 : 1
2.20	3.11	6.3 : 1
2.40	3.39	8.5 : 1
2.60	3.68	11.6 : 1
2.80	3.96	16.0 : 1
3.00	4.24	22.2 : 1
3.20	4.52	31.4 : 1
3.40	4.81	44.8 : 1
3.60	5.09	64.9 : 1
3.80	5.37	95.3 : 1
4.00	5.66	142.3 : 1
4.50	6.36	415.7 : 1
5.00	7.07	1314.8 : 1
5.50	7.78	4999.0 : 1
6.00	8.48	16665.7 : 1

*The values in this column are obtained by multiplying the values in the first column by $\sqrt{2}$.

Pearson's Tables. These tables are of general application and enable us to determine whether a deviation obtained may be due to chance or due to the nature of the experiment. The odds indicate how frequently we may expect to obtain a deviation due to chance variation as great as the one found.

It is convenient to have some standard of odds that may be accepted as indicating significance, and the standard of 30 to 1 has been generally used. It may be of interest here to determine the $\frac{D}{P.E.}$ ratio that leads to odds of 30 to 1. With odds of 30 to 1 the entire curve must be divided into $30+1$, or 31, equal parts. We then obtain 1/31 of 100000, or 3226. Subtracting 3226 from the total area, 100000, we have 96774 as the part of the area included on both sides of the mean within the limits indicated. Referring to Table VI and interpolating we find for an area of one-half 96774 that $\frac{x}{\sigma}=2.141$. For the $\frac{D}{P.E.}$ ratio, 2.141 must be divided by .6745, giving 3.174, which is the ratio of the deviation to its probable error for odds of 30 to 1. It is commonly stated that for odds of 30 to 1 the ratio of the deviation to its probable error is 3.2. In reality the ratio of 3.2 leads to odds a little more than 30 to 1, as is seen from Table 74.

Probable Errors of Different Constants. We have seen that many factors operate to affect constants that may be determined from samples drawn from a population. As a result, if several samples were drawn from a large population and a constant such as the means of the samples was determined, there would be small differences between the means. Therefore the mean or other constant that may be calculated from a sample does not represent the true value for all of the possible individuals that might have been studied. It is a true value for the sample under observation and if the sample is fairly representative of the whole population this value approaches what we may call the most probable value, and the amount it may deviate from the true value is determined by the size of the probable error. As we have seen, the size of the probable error depends on the variability of the material and the number of observations that have been studied, and we refer to the errors in such cases as being errors due to sampling.

The various constants which have been discussed in the previous chapters, as well as those that may be discussed later, are subject to these errors. The probable error of the mean has already been considered and the formula for determining it has been given. We may, however, cite another example. The data in Table 30, Chapter V, are the weights of 400 soy bean plots, and from these data the mean, standard deviation, and coefficient of variability have been determined. These constants are

$$M = 275.4375$$
$$\sigma = 49.274$$
$$C = 17.889$$

The mean of 275.4375 is the true mean for these 400 plots, but it does not represent the true value for all possible plots that might have been studied. The standard error (standard deviation) for this series of plots is 49.274. Substituting this value and the value for N in the formula for the probable error of the mean

$$P.E._M = \pm .6745 \frac{\sigma}{\sqrt{N}}$$

we have

$$P.E._M = \pm .6745 \frac{49.274}{\sqrt{400}} = \pm 1.6618$$

This probable error enables us to establish limits on each side of the mean within which it is an equal chance that the true mean will fall. In discussing the mean of these plots it is better to attach the probable error and read the mean as 275.4375 ± 1.6618. That is, to give a better idea of the most probable value of the mean it should be accompanied by its probable error. In calculating probable errors they should be carried to the same number of decimals as are the constants for which they are determined, but in some cases additional decimals will be needed for accuracy.

The probable error of the standard deviation is determined from the formula

$$P.E._\sigma = \pm .6745 \frac{\sigma}{\sqrt{2N}}$$

or it may be determined from the relation

$$P.E._\sigma = \frac{P.E._M}{\sqrt{2}}$$

Substituting in the formula the standard deviation, 49.274, and the value of N, 400, from Table 30, we have

$$P.\,E._{\sigma}=\pm.6745\frac{49.274}{\sqrt{800}}=\pm1.175$$

The probable error of the standard deviation is interpreted similarly to the probable error of the mean. That is, in this particular case the chances are equal that the true value of the standard deviation lies within the limits of 49.274±1.175. When discussing the standard deviation we should therefore accompany it by its probable error, as is done with the mean.

The probable error of the coefficient of variability, when the coefficient of variability is less than 10, is obtained from the formula

$$P\;E._{C}=\pm.6745\frac{C}{\sqrt{2N}}$$

When the coefficient of variability is more than 10 the formula is

$$P\;E._{C}=\pm.6745\frac{C}{\sqrt{2N}}\left[1+2\left(\frac{C}{100}\right)^{2}\right]^{\frac{1}{2}}$$

Substituting the value fo the coefficient of variability and N from Table 30 in the first formula for the probable error, we have

$$P.\,E._{C}=\pm.6745\frac{17.889}{\sqrt{800}}=\pm.427$$

Since in this case the coefficient of variability is more than 10, we should use the complete formula for the probable error of the coefficient of variability. Substituting the values in this formula, we have

$$P.\,E._{C}=\pm.6745\frac{17.889}{\sqrt{800}}\left[1+2\left(\frac{17.889}{100}\right)^{2}\right]^{\frac{1}{2}}=.427\times1.032=\pm.441$$

The difference in the probable error as calculated by the two formulas is .014, and the correct value is given by the complete formula.

The probable error of the coefficient of correlation is given by the formula

$$P.\,E._{r}=\pm.6745\frac{(1-r^{2})}{\sqrt{N}}$$

This same formula holds for approximate values for the probable error of the correlation ratio. Applying in this formula for *P.E.*, the values obtained in Table 36, Chapter VI, we have

$$P.E._r = \pm .6745 \frac{1-(.217)}{\sqrt{400}} = \pm .032$$

This indicates that if all possible material of this kind could be studied there is an even chance that the true value for the coefficient of correlation lies within the limits of .217±.032. We may also interpret this to mean that if we continued to make up a series of correlation tables from similar data we would expect one-half of the correlation coefficients to fall within the limits of .217±.032.

When correlations based on a large number of observations are determined, the probable error or the standard error may be used with little danger of misinterpreting the results. There are, however, many problems in which we may be interested but where the number of observations is limited to small numbers, and in such cases the standard error or the probable error is of doubtful value. Fisher has studied this problem and states: "It is necessary to warn the student emphatically against the misleading character of the standard error of the correlation coefficient deduced from a small sample, because the principal utility of the correlation coefficient lies in its application to subjects of which little is known, and upon which the data are relatively scanty."

As an aid in interpreting correlation Fisher has prepared a table of probability values. Snedecor has modified and extended this table, and his modified table appears as Table XI in the Appendix. The first column in this table refers to the degrees of freedom, which are to be discussed more fully in Chapter XIII. It is sufficient to state here that for simple correlation the degrees of freedom are taken as two less than the number of individuals in the population.

There are two sets of numbers in the table, one in light-face type and the other in dark-face type, and these numbers may be interpreted as follows. The numbers in light-face type indicate the least significant values of *r* for the particular degrees of freedom, and the numbers in dark-face type indicate the values of r that would be considered highly significant.

At the top of the table the number of variables are indicated, and these are used in the following way. In simple correlation we are studying the relation between two characters, or two variables, and for simple correlation we use the values in the second column of the table, indicated as 2 variables. This column is also used when interpreting partial correlation coefficients by considering the appropriate degrees of freedom. The degrees of freedom for partial correlation will be the total number of observations minus the number of variables concerned in the partial correlation. With multiple correlation we usually are concerned with several variables, and for interpreting multiple correlation coefficients we use the column headed by the number of variables that comes nearest to the number used in the study. For example, if six variables are concerned we would use the values in column 6 opposite the appropriate degrees of freedom.

We may now apply the values in this table by using the correlation coefficient, .217, from Table 36 in Chapter VI and interpreting its significance. The number of individuals in the population is 400, and since we are considering two variables the degrees of freedom are therefore $400-2$, or 398. In the first column of the table we find 400 for the number nearest to these degrees of freedom, and in the second column we find the value for least significance for 400 degrees of freedom to be .098, and the number for the highly significant value to be .128. These are the values we may expect due to chance if the characters are not correlated. The value obtained for the correlation coefficient, .217, is much higher than the highest value, .128, and we would therefore conclude that the correlation is highly significant and that correlation does exist between the characters.

For the interpretation of the results obtained from multiple correlation the values in the table are used in the following way. With the data from Table 57 in Chapter VIII the multiple correlation was obtained for the effect of the three variables, A, B, and C, on a fourth variable Y, and was found to be .866. There are 25 individuals in the population, and since we are considering 4 variables the degrees of freedom will be $25-4$, or 21. In the table opposite 21 degrees of freedom and under the column headed 4 variables, we find the lower value to be .552 and the higher value to be .641. Since the value .866 is much higher than these, it is evident that the coefficient of multiple correlation is very significant.

These examples illustrate how the values in this table are to be used in interpreting correlation, and it is recommended that correlation coefficients be interpreted on the basis of these values. As already indicated, when the number of individuals is large the standard error or probable error may be used, but when the number in the population is small it is better to interpret the results on the basis of the values in Table XI than to use the standard error or the probable error.

In discussing the correlation ratio a method for comparing the difference between the correlation ratio and the correlation coefficient, to determine whether there was any significant deviation from linearity, was given. This depends on the difference between the squares of η and r, and the significance of this difference is obtained from the following formula:

$$\sigma_{\eta^2-r^2}=2\sqrt{\frac{\eta^2-r^2}{N}}$$

This is the standard deviation or standard error of the difference between the squares of the two measures of correlation. The standard error is multiplied by the constant .6745 to convert it to the probable error. The complete formula for obtaining the probable error of η^2-r^2 is

$$P.E._{\eta^2-r^2}=\pm.6745\times2\sqrt{\frac{\eta^2-r^2}{N}}\sqrt{(1-\eta^2)^2-(1-r^2)^2+1}$$

Using this complete formula and substituting the values $\eta_{yx}=$ .904 and $r=.865$ from the data in Table 48, Chapter VII, we have

$$P.E._{\eta^2-r^2}=\pm.0202$$

The standard error obtained by the shorter formula is .0304, and converting this to the probable error we have .0205. Thus we see that the difference between the probable errors obtained by the two formulas is very slight, and read to three decimals there is no difference. As a rough approximation the value of $\sqrt{(1-\eta^2)^2-(1-r^2)^2+1}$ may be taken as unity and the formula reduces to

$$P.E._{\eta^2-r^2}=\pm .6745\times 2\sqrt{\frac{\eta^2-r^2}{N}}$$

The formulas for determining the probable errors of various constants are given in Table 75, pages 244, 245. It is unnecessary to show how the probable errors of all of these various constants are obtained, as the steps indicated by the formulas are simple. It is important, however, to call attention to one very useful application of the probable error in the comparison of results. This is the application of the probable error of the sum or difference.

When the sum or difference of two means, standard deviations, or other constants, is obtained, the significance of the sum or difference depends on the ratio to its probable error. That is, when a sum or difference is obtained the probable error of the sum or difference should also be determined. For example, if it is desired to obtain the difference between two independent means, M_1 and M_2, with their probable errors, E_1 and E_2, the difference between the means is obtained in the usual manner and the probable error of the difference is determined from the following formula

$$P.E._{\text{Difference}}=\sqrt{E_1^2+E_2^2}$$

In the case of the sum of the two means, M_1+M_2, the probable error of the sum is the same as the probable error of the difference. The application of this formula may now be illustrated.

Suppose two varieties of grain have been tested similarly and the average yields as obtained from ten plots are 55 ± 2.0 bushels and 50 ± 1.5 bushels. What is the difference between these two varieties, and is it significant? We have

$$55\pm 2.0$$

$$50\pm 1.5$$

Difference $\quad 5\pm\sqrt{(2.0)^2+(1.5)^2}=5\pm\sqrt{6.25}$

$$=5\pm 2.5$$

TABLE

FORMULAS FOR CALCULATING PROBABLE ERRORS

CONSTANT OR VALUE	PROBABLE ERROR
SINGLE OBSERVATION	$.6745\ \sigma$
MEAN	$.6745\ \frac{\sigma}{\sqrt{N}}$
MEDIAN	$.8453\ \frac{\sigma}{\sqrt{N}}$
STANDARD DEVIATION	$.6745\ \frac{\sigma}{\sqrt{2N}}$
QUARTILE (NORMAL DISTRIBUTION) . . .	$.9191\ \frac{\sigma}{\sqrt{N}}$
SEMI-INTERQUARTILE RANGE (NORMAL DISTRIBUTION)	$.5306\ \frac{\sigma}{\sqrt{N}}$
COEFFICIENT OF VARIABILITY	$.6745\ \frac{C}{\sqrt{2N}}\ \left[1+2\left(\frac{C}{100}\right)^2\right]^{\frac{1}{2}}$
	$.6745\ \frac{C}{\sqrt{2N}}$ (an approximation when C is not over 10)
COEFFICIENT OF CORRELATION	$.6745\ \frac{(1-r^2)}{\sqrt{N}}$
RANK CORRELATION	$.7063\ \frac{(1-r^2)}{\sqrt{N}}$
CORRELATION RATIO	$.6745\ \frac{(1-\eta^2)}{\sqrt{N}}$ (approximately)
REGRESSION COEFFICIENT . . . . β_{xy}	$.6745\ \frac{\sigma_x}{\sigma_y}\sqrt{\frac{1-r^2}{N}}$

The difference divided by its probable error, 5/2.5, is 2.0. From what has already been said we know that the possibility of deviations occurring as great or greater than twice the probable error due to chance alone are about 4.64 to 1, or about 9 to 2. That is, out of every 11 trials we would expect such deviations to occur twice due to chance alone, and we would conclude that the difference is not statistically significant.

75

OF A NUMBER OF USEFUL CONSTANTS

CONSTANT OR VALUE	PROBABLE ERROR
TO MEASURE VARIABILITY OF $P.E._s$ . .	$.4769\ \frac{P.E._s}{\sqrt{N}}$ ($\sqrt{N-1}$ in place of $\sqrt{N}$ for small numbers)
SUM OR DIFFERENCE ($A \pm a$ AND $B \pm b$). .	$\sqrt{a^2+b^2}$
MEAN OF A SERIES OF MEANS, EACH HAVING A PROBABLE ERROR.	
$M=\dfrac{\frac{A}{a^2}+\frac{B}{b^2}+\frac{C}{c^2}\cdots}{\frac{1}{a^2}+\frac{1}{b^2}+\frac{1}{c^2}\cdots}$	$\dfrac{1}{\sqrt{\frac{1}{a^2}+\frac{1}{b^2}+\frac{1}{c^2}+\cdots}}$
A SECOND METHOD FOR THE MEAN OF A SERIES OF MEANS, EACH HAVING A PROBABLE ERROR	
$M=\dfrac{\Sigma(M_1+M_2+M_3\cdots+M_n)}{N}$	$\dfrac{\sqrt{E_1^2+E_2^2+E_3^2\cdots+E_n^2}}{N}$
PRODUCT ($A \pm a \times B \pm b$).	$\sqrt{(Ab)^2+(Ba)^2}$
PRODUCT OF 3 FACTORS ($A \pm a$, $B \pm b$, $C \pm c$)	$\sqrt{(BCa)^2+(ACb)^2+(ABc)^2}$
QUOTIENT $\left(\dfrac{B \pm b}{A \pm a}\right)$.	$\dfrac{\sqrt{\left(\frac{Ba}{A}\right)^2+b^2}}{A}$

Again, let us suppose two standard deviations of the values $6.950 \pm .166$ and $5.405 \pm .129$. What is the probability that such standard deviations are obtained from the same material? We have

$$
\begin{array}{ll}
 & 6.950 \pm .166 \\
 & 5.405 \pm .129 \\
\text{Difference} & 1.545 \pm \sqrt{(.166)^2+(.129)^2} \\
 & = 1.545 \pm .210
\end{array}
$$

The deviation is over seven times its probable error, and the probability against obtaining such a difference due to chance alone is over 427000 to 1, as shown by the table of odds (Table VII) in the Appendix. From this we conclude that the two standard deviations must have arisen from samples of different material and are not different samples of the same material.

Applying this method of comparison to two coefficients of correlation obtained with the same characters, we have the following result:

$$
\begin{array}{ll}
 & .769 \pm .012 \\
 & .680 \pm .018 \\
\text{Difference} & .089 \pm \sqrt{(.012)^2 + (.018)^2} \\
 & = .089 \pm .022
\end{array}
$$

The difference divided by its probable error gives 4.05. Again referring to the table of odds in the Appendix it is seen that the odds against such a deviation being due to chance are about 158 to 1. It is reasonably safe to assume that the correlation coefficients were not determined from two samples of the same material.

These illustrations show the method of comparing constants and the importance of the application of the probable error of the difference in interpreting these comparisons.

The formula for the probable error of a sum may be extended to include several values, such as several means or other constants. For example, suppose that we had a series of means with their accompanying probable errors. We may obtain the sum and the probable error of the sum as follows:

	Mean	*P. E.*	(*P. E.*²)
	43.9	1.25	1.5625
	36.5	1.06	1.1236
	36.9	1.25	1.5625
	51.9	.51	.2601
	41.2	.72	.5184
Sum	210.4	$P.E._{\text{Sum}}=\sqrt{5.0271}$	
			$=2.24$

The sum is 210.4 and the probable error of the sum, obtained by determining the square root of the sum of the squares of the several probable errors, is 2.24. Now if the average value with its accompanying probable error is desired, it is obtained by dividing the sum and the probable error of the sum by the number of means. Dividing by 5 in this case, we have an average of 42.08±.45. For the probable error of a general average from a series of constants with their probable errors we determine

$$\frac{\sqrt{E_1^2+E_2^2+E_3^2\ldots+E_n^2}}{N}$$

When using this formula it is presumed that the several means and their probable errors are strictly comparable, that is they have been obtained under the same conditions. Another formula is available for determining an average and its probable error from a series of values, in which A, B, and C represent a series of means, and a, b, and c their respective probable errors.

$$\text{Average} = \frac{\frac{A}{a^2}+\frac{B}{b^2}+\frac{C}{c^2}\ldots}{\frac{1}{a^2}+\frac{1}{b^2}+\frac{1}{c^2}\ldots}$$

$$P.E._{\text{Average}} = \frac{1}{\sqrt{\frac{1}{a^2}+\frac{1}{b^2}+\frac{1}{c^2}\ldots}}$$

This formula gives a value for the mean or average which is influenced by the probable errors of the individual means. By substituting the means and probable errors from the data just used in determining the average and the probable error of the average, we find that the mean and its probable error is 45.72±.35. The difference in the means as obtained by the two formulas is due to the fact that in the second formula a mean with a low probable error has more influence than a mean with a high probable error. When a mean is higher than the average and at the same time has a low probable error, as in the present case where a mean of 51.9 has a probable error of±.51, this high mean with its low probable error has more influence on the general mean value than has any of the other means. The average determined by the second formula may be considered as one obtained by weighting the various means in accordance with the size of their probable errors. This formula is especially useful when one is sure that greater weight should be assigned to certain values, or when one may have reason to know that certain results are more reliable than others.

The foregoing discussion and the formulas presented indicate that all constants calculated need a qualifying or limiting value to denote how reliable they are, or, in other words, how much we may depend on them. Such an indication is given by either the standard error or the probable error. It is important to keep in mind at all times that in statistical work we are dealing with variables, and that even when a constant and its probable error have been determined the values are still subject to variation. It is well to point out that even the probable error is subject to variability, and we may obtain the probable error of the probable error of a single determination. For example, for the soy bean data in Table 30 we find the probable error of a single determination to be 33.235. The probable error of this probable error is found from the formula

$$P.E.\text{ of }P.E._s = \pm .4769 \frac{P.E._s}{\sqrt{N}}$$

Substituting the necessary values in this formula we have

$$P.E.\text{ of }P.E._s = \pm .4769 \frac{33.235}{\sqrt{400}} = \pm .793$$

When the sample is small $\sqrt{N\text{-}1}$ should be used in place of $\sqrt{N}$ in this formula.

The various formulas enable us to make some prediction as to the most probable value for the kind of data with which we are dealing but we must keep in mind that in order to make this prediction the experiments we have conducted and the material we have analyzed must be a fair sample of all the possible measurements or observations. This means that when observing single objects, such as the height of men or the income for a certain type of business, we must have large enough numbers to have a representative sample, and when dealing with individual experiments it is necessary that we conduct a sufficient number of experiments so that the samples we are analyzing represent fairly what may be expected in general from similar experiments. Merely because a probable error is small we should not conclude that our data represent fairly all the possible conditions. We must be certain that the data on which the calculations have been based are representative of the entire population.

CHAPTER X

CURVE FITTING

It is often important for purposes of generalization and prediction to fit curves to data that have been observed. We have already seen in Chapter VII how simple curve fitting proceeds, and in that chapter methods for fitting the straight line and parabolas were given. The methods of curve fitting may be extended and curves fitted to various types of frequency distributions. The derivation of the formulas for the fitting of curves is beyond the scope of this text, but the application of the necessary formulas will be illustrated by fitting some of the more important types of curves.*

Each particular curve type has its own method of analysis, and in order to determine to what type a curve belongs it is necessary to consider certain constants. The constants most commonly used for denoting curve types are the β values, namely β_1 and β_2, and they are determined by first obtaining the following values or moments about the assumed mean.

$\nu_1 = \frac{\Sigma fD}{N}$ or the mean product of the frequencies times the deviation of each class measured from the assumed mean of the distribution.

$\nu_2 = \frac{\Sigma fD^2}{N}$ or the mean product of the frequencies times the square of the deviation of each class measured from the assumed mean of the distribution.

$\nu_3 = \frac{\Sigma fD^3}{N}$ or the mean product of the frequencies times the cube of the deviation of each class measured from the assumed mean of the distribution.

---

*The author is indebted to Dr. Raymond Pearl, of Johns Hopkins University, for valuable suggestions relative to curve fitting.

$\nu_4 = \frac{\Sigma fD^4}{N}$ or the mean product of the frequencies times the fourth power of the deviation of each class measured from the assumed mean of the distribution.

In determining these deviations the usual method is followed. The mid-point of the class represents the class, and the deviation of each class in turn from the assumed mean is obtained and the steps carried forward as indicated. The calculations may be made on the unity-step basis, remembering to make the necessary correction for the class interval when determining the mean, mode, or any constant that is expressed in the original unit of measurement. It is very important to observe the signs in obtaining the sum of the products. Negative signs will appear in the calculations for the first and third moments.

When these moments have been obtained the values for μ_2, μ_3, and μ_4, which are the moments about the true mean, are calculated as follows:

$$\mu_1 = 0$$
$$\mu_2 = \nu_2 - \nu_1^2$$
$$\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3$$
$$\mu_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4$$

In distributions where a large class interval has been used, and also in cases of curves which are known to have high contact, that is where there is a tendency for the curve to tail off gradually at the extremes, it may be desirable to make certain corrections for μ_2 and μ_4. These are known as Sheppard's corrections, and by their addition the values for μ_2 and μ_4 are modified as follows:

$$\mu_2 = \nu_2 - \nu_1^2 - \tfrac{1}{12}$$
$$\mu_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4 - \tfrac{1}{2}(\nu_2 - \nu_1^2) + \tfrac{7}{240}$$

The values for β_1 and β_2 are obtained from

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

In addition to β_1 and β_2 the constants κ_1 and κ_2 may at times be needed, and they are obtained from

$$\kappa_1 = 2\beta_2 - 3\beta_1 - 6$$

$$\kappa_2 = \frac{\beta_1(\beta_2+3)^2}{4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6)} \text{ or } \frac{\beta_1(\beta_2+3)^2}{4(4\beta_2-3\beta_1)\kappa_1}$$

For certain curve types it is necessary to have the theoretical mode and it is obtained from

$$\text{Mode} = \text{Mean} - d$$

when

$$d = \frac{\mu_3(\mu_4+3\mu_2^2)}{2(5\mu_2\mu_4-6\mu_3^2-9\mu_2^3)}$$

The skewness may be determined from the formula

$$\text{Skewness} = \frac{\sqrt{\beta_1}\,(\beta_2+3)}{2(5\beta_2-6\beta_1-9)}$$

The skewness may be defined as a measure of asymmetry, or departure of a frequency distribution from the symmetrical or bell-shaped type of distribution. Previously we have discussed certain constants of position, namely, the mode, median, and mean, and the skewness may be considered as the ratio of the difference between the mean and the mode to the standard deviation of a frequency distribution, or

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

In order that there may be no mistake in the sign of the skewness, the general formula given above for obtaining skewness in terms of the β values is the more reliable.

After the β and κ values have been calculated we may judge to what type a curve belongs from the chart on page 252. Only the formulas and criteria for the curves presented here are given. It is more convenient if Pearson's Tables are available to refer to the complete chart from which the curve type may be determined rather quickly from the β values. It may be stated that if much curve fitting is to be done Pearson's Tables will be found indispensable, as many of the required values are tabled and time will be saved in making the various calculations.

CHART SHOWING EQUATION, ORIGIN, AND CRITERION VALUES FOR VARIOUS CURVE TYPES

NUMBER OF TYPE	EQUATION	ORIGIN	CRITERION
MAIN TYPES			
I	$Y=Y_0(1+x/a_1)^{m_1}(1-x/a_2)^{m_2}$	Mode	κ_2 negative
IV	$Y=\frac{Y_0}{[1+(x/a)^2]^m} e^{-v \tan^{-1} x/a}$	Mean $+(va/r$ times the class interval)	$\kappa_2>0$ and <1
VI	$Y=Y_0(x-a)^{q_2}x^{-q_1}$	Mean $-\mu_1'$	$\kappa_2>1$
TRANSITION TYPES			
Normal Curve	$Y=\frac{N}{\sigma\sqrt{2\pi}} e^{-(x^2/2\sigma^2)}$	Mode (=Mean)	$\kappa_2=0,\ \beta_1=0,\ \beta_2=3$
II	$Y=Y_0(1-x^2/a^2)^m$	Mode (=Mean)	$\kappa_2=0,\ \beta_1=0,\ \beta_2<3$
III	$Y=Y_0(1+x/a)^{\gamma a} e^{-\gamma x}$	Mode	$2\beta_2=6+3\beta_1$
V	$Y=Y_0 x^{-p} e^{-\gamma/x}$	End of curve	$\kappa_2=1$

As already indicated, the process of curve fitting enables us to predict the expected values if we could observe or measure all of the individuals in a group rather than only a small sample, or to predict what may be expected in general from other samples of the same kind. The observed frequency distribution is usually not a very smooth distribution since there may be various unexpected deviations due to errors of sampling, and so it is desirable to have a smoothed curve for generalizing or predicting.

As indicated above, the analysis of curve fitting has resulted in developing a number of curve types. It is unnecessary to illustrate the methods used in calculating all of these types, but a few of the more commonly used will be fitted. The curve types are now divided into two main groups, the main types and the transition types. We will fit examples of each type, but for more extensive work it will be necessary to refer to special treatment of the subject, on which considerable literature is available. With each type a comparison is made between the observed and calculated values.

We will consider first a curve of Type I, which is one of the three main types. The necessary steps for fitting a curve of Type I follow.

CURVE TYPE I

Curves of Type I have the range limited in both directions and are asymmetrical or skew. The equation is

$$Y = Y_o \left(1+\frac{x}{a_1}\right)^{m_1} \left(1-\frac{x}{a_2}\right)^{m_2}$$

The constants needed are as follows:

$$r = \frac{6(\beta_2-\beta_1-1)}{-\kappa_1}$$

$$w = 4+\left[\frac{(\beta_1)}{4}\,\frac{(r+2)^2}{r+1}\right]$$

$$\epsilon = \frac{r^2}{w}$$

$$b^2 = \mu_2 w(r+1)$$

$$b = \sqrt{b^2} = \text{the range}$$

$$m'_{1,2} = \frac{r \pm \sqrt{r^2-4\epsilon}}{2}$$

$$m'_1 - 1 = m_1, \text{ and } m'_2 - 1 = m_2$$

To determine which of the roots of the quadratic is to be taken as m_1 and which as m_2, the following general rule may be applied.

When μ_3 is positive, $m_2 > m_1$, when μ_3 is negative, $m_1 > m_2$

in which $>$ is to be taken as signifying *absolutely* greater than, not merely numerically greater than. Thus, suppose μ_3 to be positive and both m_1 and m_2 negative. Then m_1 will be numerically greater than m_2, but absolutely smaller.

$$a_2 = \frac{b}{\frac{m_1}{m_2}+1} = \text{distance from mode to upper } (+) \text{ end of range}$$

$$a_1 = b - a_2 = \text{distance from mode to lower } (-) \text{ end of range}$$

$$Y_o = \frac{N}{b}\,\frac{m_1^{m_1} m_2^{m_2}}{(m_1+m_2)^{m_1+m_2}}\,\frac{\Gamma(m_1+m_2+2)}{\Gamma(m_1+1)\Gamma(m_2+1)}$$

For the calculation of Y_0 by the use of logarithms, the Γ values may be determined from Table XXXI in *Pearson's Tables for Statisticians and Biometricians.* When n is numerically greater than say 12, an approximation to Γ may be used. One such approximation is

$$\Gamma(n+1)=\sqrt{2\pi}\left(\frac{\sqrt{n^2+n+1/6}}{e}\right)^{n+\frac{1}{2}},$$ where $n+1$ is the value for which Γ is to be determined.

From the value of Y_0 we obtain

$$\kappa=\frac{Y_0}{a_1^{m_1}\ a_2^{m_2}}$$

To determine the mode we obtain

$$d=\frac{\mu_3(\mu_4+3\mu_2^2)}{2(5\mu_2\mu_4-6\mu_3^2-9\mu_2^3)}$$

$$\text{Mode}=\text{Mean}-d$$

The following data on total yield of plant in grams for oats are used to illustrate the calculation of a Type I curve.

Total Yield of Plant in Grams	f
0.00–0.99	3
1.00–1.99	50
2.00–2.99	106
3.00–3.99	109
4.00–4.99	80
5.00–5.99	42
6.00–6.99	7
7.00–7.99	2
8.00–8.99	1
	Total 400

The following products are needed for determining the moments.

V	f	D	fD	fD^2	fD^3	fD^4
.5	3	−3	− 9	27	− 81	243
1.5	50	−2	−100	200	−400	800
2.5	106	−1	−106	106	−106	106
3.5	109	0				
4.5	80	1	80	80	80	80
5.5	42	2	84	168	336	672
6.5	7	3	21	63	189	567
7.5	2	4	8	32	128	512
8.5	1	5	5	25	125	625
	N = 400		Σ = −17	701	271	3605

From these summation values we have

$$\nu_1 = \frac{-17}{400} = -.0425$$

$$\nu_2 = \frac{701}{400} = 1.7525$$

$$\nu_3 = \frac{271}{400} = .6775$$

$$\nu_4 = \frac{3605}{400} = 9.0125$$

Substituting these values in the equations for μ_2, μ_3, and μ_4, we have

$$\mu_2 = 1.7525 - (-.0425)^2 = 1.7507$$

$$\mu_3 = .6775 - 3(-.0425)(1.7525) + 2(-.0425)^3 = .9008$$

$$\mu_4 = 9.0125 - 4(-.0425)(.6775) + 6(-.0425)^2(1.7525) - 3(-.0425)^4$$
$$= 9.1469$$

Substituting these values in the formulas for β_1 and β_2 we have

$$\beta_1 = \frac{(.9008)^2}{(1.7507)^3} = .1512$$

$$\beta_2 = \frac{9.1469}{(1.7507)^2} = 2.9843$$

For κ_1 and κ_2 we have

$$\kappa_1 = 2(2.9843) - 3(.1512) - 6 = -.4850$$

$$\kappa_2 = \frac{.1512\ (2.9843+3)^2}{4[4(2.9843) - 3(.1512)]\ [2(2.9843) - 3(.1512) - 6]} = -.2430$$

From the necessary constants we obtain

$$r = \frac{6(2.9843 - .1512 - 1)}{-(-.4850)} = 22.6775$$

$$w = 4 + \left[\frac{.1512}{4}\ \frac{(22.6775+2)^2}{22.6775+1}\right] = 4.9722$$

$$\epsilon = \frac{(22.6775)^2}{4.9722} = 103.4289$$

$$b^2 = 1.7507 \times 4.9722 \times (22.6775+1) = 206.1079$$

$$b = \sqrt{206.1079} = 14.3565$$

$$m'_{1,2} = \frac{22.6775 \pm \sqrt{(22.6775)^2 - 413.7156}}{2}$$

$$= \frac{22.6775 \pm 10.0276}{2} \quad \text{and}$$

$$\frac{22.6775 + 10.0276}{2} = 16.3525$$

$$\frac{22.6775-10.0276}{2}=6.3249$$

Following the general rule we determine which root is m_1 and which is m_2.

When μ_3 is positive, $m_2>m_1$, when μ_3 is negative, $m_1>m_2$

In the case at hand μ_3 is positive, therefore $m_2>m_1$, and

$$m'_2=16.3525$$
$$m'_1=6.3249$$
$$m_1=6.3249-1=5.3249$$
$$m_2=16.3525-1=15.3525$$
$$a_2=\frac{14.3565}{\frac{5.3249}{15.3525}+1}=10.6597$$
$$a_1=14.3565-10.6597=3.6968$$

$$Y_o=\frac{400}{14.3565}\,\frac{\left(5.3249^{5.3249}\right)\left(15.3525^{15.3525}\right)}{(5.3249+15.3525)^{5.3249+15.3525}}\,\frac{\Gamma(5.3249+15.3525+2)}{\Gamma(5.3249+1)\Gamma(15.3525+1)}$$

Y_o is calculated by logarithms, as follows:

$\log 400 = 2.6020600$

$\log 5.3249 = .7263115$, and $.7263115 \times 5.3249 = 3.8675361$

$\log 15.3525 = 1\ 1861791$, and $1.1861791 \times 15.3525 = 18.2108146$

$\Gamma(5.3249+15.3525+2)=\Gamma 22.6774$

To obtain $\Gamma 22.6774$ we use the approximation formula cited above and proceed as follows. For $\Gamma 22.6774$ we take n as 21.6774, since in obtaining the Γ function we use 1 less than the number. Then

$$\sqrt{n^2+n+1/6}=\sqrt{469.9097+21.6774+.1667}=\sqrt{491.7538}$$

$$\log \sqrt{491.7538}=1.3458738$$

$$\log\left(\frac{\sqrt{n^2+n+1/6}}{e}\right)^{n+\frac{1}{2}}=(1.3458738-.4342945)\times 22.1774=20.2164588$$

$$\log \sqrt{2\pi}=.399090$$

$$\log \sqrt{2\pi}\left(\frac{\sqrt{n^2+n+1/6}}{e}\right)^{n+\frac{1}{2}}=.399090+20.2164588=20.6155488$$

$$\log \Gamma(5.3249+15.3525+2)=20.6155488$$

$$\log 14.3565= 1.1570485$$

$$\log (5.3249+15.3525)^{5.3249+15.3525}=\log 20.6774\times 20.6774$$

$$=1.3154959\times 20.6774=27.2010349$$

For $\Gamma(5.3249+1)$, since n is small, we proceed as follows:

$\Gamma 6.3249 =$ log 5.3249 or .7263115
+log 4.3249 or .6359761
+log 3.3249 or .5217786
+log 2.3249 or .3664043
+log 1.3249 or .1221831
+ $\Gamma 1.3249$ or $\bar{1}.9513421$ *(determined from Pearson's Table XXXI)**

Sum $= \Gamma 6.3249 = 2.3239957$

For $\Gamma(15.3525+1)$ we again use the approximation formula. For $\Gamma(n+1)$ for the value 16.3525, taking n as 15.3525, we have

$$\sqrt{(n^2+n+1/6)}=\sqrt{235.6993+15.3525+.1667}=\sqrt{251.2185}$$

$$\log \sqrt{251.2185}=1.2000258$$

$$\log \left(\frac{\sqrt{(n^2+n+1/6)}}{e}\right)^{n+\frac{1}{2}}=(1.2000258-.4342945)\times 15.8525=12.1387554$$

$$\log \sqrt{2\pi}=.399090$$

$$\log \sqrt{2\pi}\left(\frac{\sqrt{(n^2+n+1/6)}}{e}\right)^{n+\frac{1}{2}}=.399090+12.1387554=12.5378454$$

Combining the logarithms we have log Y_o.

$$\begin{array}{lr}
\log N & = 2.6020600\\
\log m_1^{m_1} & = 3.8675361\\
\log m_2^{m_2} & =18.2108146\\
\log \Gamma(m_1+m_2+2) & =20.6155488\\
\text{colog } b & = \bar{2}.8429515\\
\text{colog } (m_1+m_2)^{m_1+m_2} & =\overline{28}.7989651\\
\text{colog } \Gamma(m_1+1) & = \bar{3}.6760043\\
\text{colog } \Gamma(m_2+1) & =\underline{\overline{13}.4621546}\\
\log Y_o = & 2.0760350
\end{array}$$

By logarithms

$$\log \kappa=2.0760350-(\log 3.6968\times 5.3249)+(\log 10.6597\times 15.3525)$$

$$\log \kappa=\overline{17}.2739632$$

---

* If Pearson's Tables are not available the approximation formula for Γ may be used, keeping in mind that it is not so exact for small numbers as the values from Pearson's Table.

$$d = \frac{.9008\ (9.1469+9.1950)}{2(80.0675-4.8684-48.2931)} = .3070$$

$$\text{Mode} = 3.4575 - .3070 = 3.1505$$

Columns are formed and the work carried out as follows. The values in column 2 are the deviations of the ordinates from the mode, which is the origin. Completing the operations we have the calculated values in column 10, which may be compared with the observed values as given in column 11. Methods for comparing the observed with the calculated values are given in Chapter XI.

DETAILS OF DETERMINING CALCULATED Y VALUES FOR TYPE I CURVE

1	2	3	4	5	6	7	8	9	10	11
V	x	a_1+x	LOG (a_1+x)	LOG (a_1+x) TIMES m_1	a_2-x	LOG (a_2-x)	LOG (a_2-x) TIMES m_2	COLUMN 5+ COLUMN 8+ LOG κ= LOG Y	CALCULATED Y	OBSERVED Y
.5	−2.6505	1.0463	.0196562	.1046673	13.3102	1.1241846	17.2590441	.6376746	4.34	3
1.5	−1.6505	2.0463	.3109693	1.6558804	12.3102	1.0902651	16.7382949	1.6681385	46.57	50
2.5	− .6505	3.0463	.4837727	2.5760413	11.3102	1.0534703	16.1734028	2.0234073	105.54	106
3.5	.3495	4.0463	.6070581	3.2325237	10.3102	1.0132671	15.5561832	2.0626701	115.52	109
4.5	1.3495	5.0463	.7029731	3.7432615	9.3102	.9689590	14.8759430	1.8931677	78.19	80
5.5	2.3495	6.0463	.7814897	4.1613545	8.3102	.9196115	14.1183356	1.5536533	35.78	42
6.5	3.3495	7.0463	.8479611	4.5153081	7.3102	.8639293	13.2634746	1.0527459	11.29	7
7.5	4.3495	8.0463	.9055962	4.8222092	6.3102	.8000431	12.2826617	.3788341	2.39	2
8.5	5.3495	9.0463	.9564710	5.0931124	5.3102	.7251109	11.1322651	$\bar{1}$.4993407	.32	1

CURVE TYPE IV

A curve of Type IV has unlimited range. The equation is

$$Y = \frac{Y_0}{[1+(x/a)^2]^m} e^{-v \tan - x/a}$$

but, since

$$x = a \tan \theta$$

and

$$\frac{1}{[1+(x/a)^2]^m} = \cos^{2m}\theta$$

then

$$Y = Y_0 \cos^{2m}\theta \, e^{-v\theta}$$

The constants of the curve are given by the following equations

$$r = \frac{6(\beta_2 - \beta_1 - 1)}{2\beta_2 - 3\beta_1 - 6}$$

$$m = \tfrac{1}{2}(r+2)$$

$$Z = \frac{r^2}{1 - \left[\frac{\beta_1}{16} \frac{(r-2)^2}{r-1}\right]}$$

$v = \sqrt{(Z - r^2)}$ *(to be given opposite sign to that of* μ_3 *in the actual statistics)*

$$a = r\sqrt{\frac{\mu_2(r-1)}{Z}}$$

The following is a close approximation for Y_0 where $\tan \phi = \frac{v}{r}$

$$Y_0 = \frac{N}{a}\sqrt{\frac{r}{2\pi}} \frac{e^{(\cos^2 \phi)/3r - 1/12r - \phi v}}{(\cos \phi)^{r+1}}$$

If tables of the $G(r, v)$ integrals, as given in Table LIV of *Pearson's Tables for Statisticians and Biometricians* are available, the formula for Y_0 may be conveniently put in the following form:

$$Y_0 = \frac{N}{a} \frac{1}{F(r, v)}$$

Log $F(r, v)$ is tabulated for values of r from 1 to 50 and of ϕ from 0° to 45° when $\tan \phi = \frac{v}{r}$, in Table LIV of *Pearson's Tables for Statisticians and Biometricians.*

To find Y_0 take

$$\text{Tan } \phi = \frac{v}{r}$$

$$\text{Log tan } \phi = \log \frac{v}{r}$$

From log tan ϕ we can determine ϕ in degrees, minutes, and seconds, from tables of trigonometric values included with most logarithm tables. These can be reduced to degrees and fraction of a degree by multiplying the given number of minutes by 60, adding the given number of seconds to the result, thus getting the number of seconds in the fraction of a degree, and dividing the result by 3600, the number of seconds in a degree.

Assume that n is the integral part of ϕ and m is the integral part of r. Then from Pearson's Table, cited above, substitute the proper values of log $F(r, v)$ for ϕ and r in a tabulation as follows:

For ϕ (n) and r (m) we have the value designated as μ_{00}
For ϕ (n) and r $(m+1)$ we have the value designated as μ_{01}
For ϕ $(n+1)$ and r (m) we have the value designated as μ_{10}
For ϕ $(n+1)$ and r $(m+1)$ we have the value designated as μ_{11}

Write down the quantities

p = fractional part of ϕ
r' = fractional part of r
$q = 1 - p$
$s = 1 - r'$

The various values of μ from the tabulation are multiplied by the values of p, r', q, and s, in the following combinations. The sum of the products is log $F(r, v)$.

$$\begin{array}{c} q\,s\,\mu_{00} \\ q\,r'\,\mu_{01} \\ p\,s\,\mu_{10} \\ \underline{p\,r'\,\mu_{11}} \\ \text{Sum} = \text{Log } F(r, v) \end{array}$$

The accuracy of the work may be judged by remembering that the largest coefficient will belong to the term that comes nearest to the actual values of ϕ and r. Then

$$Y_o = \frac{N}{a\,F(r,v)}$$

The origin of the curve is at a distance from the mean of

$$\frac{va}{r} \text{ times the class interval}$$

or

$$\text{Origin} = \text{Mean} + \frac{va}{r} \text{ times the class interval}$$

In determining the origin, the sign of v must be observed.

$$\text{Mode} = \text{Mean} - d$$

$$d = \frac{2va}{r(r+2)} \text{ times the class interval} \quad \textit{(To be given sign of } \mu_3\textit{)}$$

To illustrate the method of calculating a Type IV curve, the following data on height of culm in oats are used.

HEIGHT OF CULM IN CENTIMETERS	f
25.00–29.99	4
30.00–34.99	7
35.00–39.99	27
40.00–44.99	76
45.00–49.99	151
50.00–54.99	237
55.00–59.99	196
60.00–64.99	83
65.00–69.99	27
70.00–74.99	15
75.00–79.99	2
TOTAL	825

Constants are obtained as follows:

$$\mu_2 = 2.3795$$
$$\mu_3 = -.2620$$
$$\mu_4 = 20.1425$$
$$\beta_1 = .0051$$

$$\beta_2 = \quad 3.5575$$
$$\kappa_2 = \quad .0035$$

From the necessary constants we have

$$r = \frac{6(3.5575 - .0051 - 1)}{1.0997} = 13.9260$$

$$m = \tfrac{1}{2}(13.9260 + 2) = 7.9630$$

$$Z = \frac{(13.9260)^2}{1 - \left[\dfrac{.0051}{16}\,\dfrac{(13.9260-2)^2}{13.9260-1}\right]} = 194.5756$$

$$v = \sqrt{194.5756 - (13.9260)^2} = .8013 \quad \textit{(given opposite sign to that of } \mu_3\textit{)}$$

$$a = 13.9260 \sqrt{\frac{2.3795\ (13.9260-1)}{194.5756}} = 5.5370$$

$$\text{Tan } \phi = \frac{.8013}{13.9260} = .0575$$

$$\text{Log tan } \phi = 8.7596678$$

$$\phi \text{ in degrees} = 3^\circ\ 17'\ 27'', \text{ or } 3.2908^\circ$$

$$n = 3$$
$$m = 13$$
$$p = \quad .2908$$
$$r' = \quad .9260$$
$$q = \quad 1 - .2908 = .7092$$
$$s = \quad 1 - .9260 = .0740$$

Substituting for ϕ and r the values of log F (r, v) as found in Pearson's Table of the G (r, v) integrals, for n and m in the various combinations we have

$$\phi\ (3)\ r\ (13), \text{ or } \mu_{00} = 9.8409592$$
$$\phi\ (3)\ r\ (14), \text{ or } \mu_{01} = 9.8260558$$
$$\phi\ (4)\ r\ (13), \text{ or } \mu_{10} = 9.8465615$$
$$\phi\ (4)\ r\ (14), \text{ or } \mu_{11} = 9.8321212$$

Substituting we have

$$q\ s\ \ \mu_{00} = (.7092)\ (.0740)\ (9.8409592) = 9.9916504$$
$$q\ r'\ \mu_{01} = (.7092)\ (.9260)\ (9.8260558) = 9.8857708$$
$$p\ s\ \ \mu_{10} = (.2908)\ (.0740)\ (9.8465615) = 9.9967011$$
$$p\ r'\ \mu_{11} = (.2908)\ (.9260)\ (9.8321212) = 9.9547902$$
$$\text{Sum} = \text{Log } F(r,v) \qquad = 9.8289125$$

Then, from the formula $Y_0 = \frac{N}{a\,F(r,v)}$ we obtain by logarithms the log Y_0 as follows:

$$\begin{aligned} \log N &= 2.9164539 \\ \text{colog } a &= \bar{1}.2567255 \\ \text{colog } F(r,v) &= \underline{.1710875} \\ \text{Log } Y_0 &= 2.3442669 \end{aligned}$$

By the general approximation formula, Y_0 is obtained as follows:

$$Y_0 = \frac{N}{a}\sqrt{\frac{r}{2\pi}}\,\frac{e^{(\cos^2\phi)/3r - 1/12r - \phi v}}{(\cos\phi)^{r+1}}$$

Using logarithms, Y_0 by this formula is calculated as follows:

$$\begin{aligned} N &= 825 \\ a &= 5.5370 \\ r &= 13.9260 \\ 2\pi &= 6.2831854 \\ e &= 2.7182818 \\ \tan\phi &= 3^\circ\ 17'\ 27'' = .0574348 \text{ (in radians)} \\ \log\cos\phi &= 9.9992833 \\ \log\cos^2\phi &= 9.9985666 \\ \cos^2\phi &= .9967 \\ v &= .8013 \\ \log r(13.9260) &= 1.1438264 \\ \log 2\pi(6.2831854) &= .7981799 \\ \log\sqrt{\frac{r}{2\pi}} &= .1728232 \\ \frac{\cos^2\phi}{3r} &= \frac{.9967}{41.7780} = .0239 \\ \frac{1}{12r} &= \frac{1}{167.1120} = .0060 \\ \phi v &= (.0574348)\,(.8013) = .0460 \\ \frac{\cos^2\phi}{3r} - \frac{1}{12r} - \phi v &= .0239 - .0060 - .0460 = -.0281 \end{aligned}$$

$$e^{-.0281} = (\log e)\,(-.0281) = (.4342945)\,(-.0281) = -.0122037 = \bar{1}.9877963$$

$$\text{colog } \cos\phi^{r+1} = (9.9992833)\,(14.9260) = 9.9893025 = .0106975$$

Summing the necessary values to obtain the logarithm for Y_0, we have

$$
\begin{aligned}
\log N &= 2.9164539\\
\operatorname{colog} a &= \bar{1}.2567255\\
\log\sqrt{\frac{r}{2\pi}} &= .1728232\\
\log e^{(\cos^2\phi)/3r-1/12r-\phi v} &= \bar{1}.9877963\\
\operatorname{colog}(\cos\phi)^{r+1} &= \underline{.0106975}\\
\operatorname{Log} Y_0 &= 2.3444964
\end{aligned}
$$

This agrees very closely with the value for log Y_0, 2.3442669, obtained by the formula

$$Y_0=\frac{N}{aF(r,v)}$$

$$\text{Origin}=52.9910+\frac{(.8013)\,(5.5370)}{13.9260}\text{ times class interval}=54.5840$$

$$d=\frac{2(.8013)\,(5.5370)}{13.9260\,(13.9260+2)}\text{ times class interval}=-.2000$$

The value for d has been given the sign of μ_3.

$$\text{Mode}=52.9910-(-.2000)=53.1910$$

Columns are formed and the work carried out as indicated, x being the deviation of the ordinate from the origin. For the values in column 3 in this particular case it was found simpler to obtain the factor $\frac{1}{(a)\,(5)}$ and multiply the several values of x by this factor for the values of x/a. The value for $Y_0=2.3442669$ was used in determining the values in column 11.

DETAILS OF DETERMINING CALCULATED Y VALUES FOR TYPE IV CURVE

1	2	3	4	5	6	7	8	9	10	11	12	
V		TAN θ $= x/a$	LOG TAN θ	$\theta°$	LOG COS θ	LOG COS θ TIMES $2m$	θ IN RADIANS	$v\theta$	LOG $e^{-v\theta}$	COLUMN 7+ COLUMN 10+ LOG Y_0= LOG Y	CALCULATED Y	OBSERVED Y
27.5	-27.0840	-.9777	9.9902056	44° 21′ 14″	9.8543277	$\bar{3}.6800230$	-.7741214	-.6203035	.2693944	.2936843	1.97	4
32.5	-22.0840	-.7972	9.9015673	38° 33′ 43″	9.8931705	$\bar{2}.2986334$	-.6730329	-.5393013	.2342156	.8771159	7.54	7
37.5	-17.0840	-.6167	9.7900739	31° 39′ 44″	9.9300099	$\bar{2}.8853377$	-.5526100	-.4428064	.1923084	1.4219130	26.42	27
42.5	-12.0840	-.4362	9.6396857	23° 34′ 1″	9.9621768	$\bar{1}.3976277$	-.4113208	-.3295914	.1431397	1.8850343	76.74	76
47.5	- 7.0840	-.2557	9.4077307	14° 20′ 36″	9.9862470	$\bar{1}.7809697$	-.2503384	-.2005962	.0871178	2.2123544	163.06	151
52.5	- 2.0840	-.0752	8.8762178	4° 18′ 2″	9.9987755	$\bar{1}.9804986$	-.0750589	-.0601447	.0261205	2.3508860	224.33	237
57.5	2.9160	.1053	9.0224284	6° 0′ 40″	9.9976055	$\bar{1}.9618652$	.1049137	.0840673	$\bar{1}.9634900$	2.2696221	186.05	196
62.5	7.9160	.2858	9.4560622	15° 57′ 0″	9.9829501	$\bar{1}.7284633$	.2783800	.2230659	$\bar{1}.9031237$	1.9758539	94.59	83
67.5	12.9160	.4063	9.6086654	24° 59′ 59″	9.9572767	$\bar{1}.3195887$	.4363275	.3496292	$\bar{1}.8481580$	1.5120136	32.51	27
72.5	17.9160	.6468	9.8107700	32° 53′ 41″	9.9241085	$\bar{2}.7913520$	.5741212	.4600433	$\bar{1}.8002057$	.9358246	8.63	15
77.5	22.9160	.8273	9.9176630	39° 36′ 3″	9.8867749	$\bar{2}.1967771$	.6911649	.5538304	$\bar{1}.7594745$	.3005185	2.00	2

CURVE TYPE VI

A curve of Type VI has the range limited on one side. The equation is

$$Y = Y_o \, (x-a)^{q_2} \, x^{-q_1}$$

The following constants are needed:

$$r = \frac{6(\beta_2 - \beta_1 - 1)}{-\kappa_1}$$

$$w = 4 + \left[\frac{\beta_1}{4} \, \frac{(r+2)^2}{r+1} \right]$$

$$\epsilon = \frac{r^2}{w}$$

These three constants are determined as in Type I. The values for r and ϵ will both be negative if the curve is Type VI.

$1-q_1$ and q_2+1 are the roots of $Z^2 - rZ + \epsilon = 0$

Calculate

$$Z_1 Z_2 = \frac{r \pm \sqrt{r^2 - 4\epsilon}}{2}$$

Then

$$q_1 = 1 - Z_1$$
$$q_2 = Z_2 - 1$$

In taking the roots of the quadratic it should be remembered that in a curve of this type ϵ must be negative, and the following relations necessarily follow.

The value for r must be negative. But $r = -q_1 + q_2 + 2$
Therefore q_1 must be $> q_2 + 2$.
Then Z_1 and Z_2 must be taken so as to fulfill this requirement

$$a^2 = \frac{\mu_2 \, r^2 \, (r+1)}{(1-q_1)\,(1+q_2)}$$

$a = \sqrt{a^2}$ *(distance from origin to range end)*

$$Y_o = \frac{N(a^{q_1 - q_2 - 1})\Gamma q_1}{\Gamma(q_1 - q_2 - 1)\Gamma(q_2 + 1)}$$

The origin is distant from the mean by a quantity

$$\mu_1' = \frac{a(q_1-1)}{q_1-q_2-2}$$

in which a is to be given the sign of μ_3 in the actual statistics.

Origin = Mean − μ_1', having regard to the sign of μ_1'

$$\text{Mode} = \text{Mean} - \tfrac{1}{2}\left[\frac{\mu_3}{\mu_2}\,\frac{r+2}{r-2}\right]$$

The following data on the number of culms in oats are used to illustrate the calculation of a Type VI curve.

Number of Culms	f
1	1
2	57
3	184
4	111
5	33
6	11
7	2
8	1
Total	400

The following constants are obtained:

$$\mu_2 = .9819$$
$$\mu_3 = .8733$$
$$\mu_4 = 4.3992$$
$$\beta_1 = .8057$$
$$\beta_2 = 4.5630$$
$$\kappa_1 = .7089$$
$$\kappa_2 = 1.0264$$

From the necessary constants we have

$$r = \frac{6(4.5630 - .8057 - 1)}{-.7089} = -23.3373$$

$$w = 4 + \left[.2014\frac{(-23.3373+2)^2}{-23.3373+1}\right] = -.1050$$

$$\epsilon = \frac{(-23.3373)^2}{-.1050} = -5186.9486$$

$$Z_1 Z_2 = \frac{-23.3373 \pm \sqrt{(-23.3373)^2 - (-20747.7944)}}{2}$$

$$Z_1 = -84.6282$$

$$Z_2 = 61.2909$$

$$q_1 = 1 - (-84.6282) = 85.6282$$

$$q_2 = 61.2909 - 1 = 60.2909$$

$$a^2 = \frac{.9819\ (-23.3373)^2\ (-23.3373+1)}{(1-85.6282)\ (1+60.2909)} = 2.3030$$

$$a = \sqrt{2.3030} = 1.5176 \quad \textit{(given the sign of } \mu_3\textit{)}$$

$$Y_o = \frac{400\left(1.5176^{85.6282-60.2909-1}\right)\Gamma 85.6282}{\Gamma(85.6282-60.2909-1)\Gamma(60.2909+1)}$$

Calculating Y_0 by logarithms we have

$$\log Y_o = 29.4279775$$

$$\mu_1' = \frac{1.5176\ (85.6282-1)}{(85.6282-60.2909-2)} = 5.5033$$

$$\text{Origin} = 3.4100 - 5.5033 = -2.0933$$

$$\text{Mode} = 3.4100 - \tfrac{1}{2}\left[\frac{.8733}{.9819}\ \frac{-23.3373+2}{-23.3373-2}\right] = 3.0355$$

Columns are arranged and the work carried out as follows, x being the deviation of the ordinate from the origin.

DETAILS OF DETERMINING CALCULATED Y VALUES FOR TYPE VI CURVE

1	2	3	4	5	6	7	8	9	
V	x	$x-a$	LOG x	LOG $x-a$	(LOG $x-a$)q_2	$\left(\frac{1}{\text{LOG } x}\right)q_1$	COLUMN 6+ COLUMN 7+ LOG Y_0= LOG Y	CALCULATED Y	OBSERVED Y
1	3.0933	1.5757	.4904220	.1974735	11.9058550	$\overline{42}.0060469$	$\bar{1}.3398794$	.22	1
2	4.0933	2.5757	.6120736	.4108953	24.7732474	$\overline{58}.5892394$	1.7904643	61.73	57
3	5.0933	3.5757	.7069993	.5533611	33.3626387	$\overline{61}.4609225$	2.2515387	178.46	184
4	6.0933	4.5757	.7848526	.6604575	39.8195771	$\overline{68}.7944846$	2.0420392	110.16	111
5	7.0933	5.5757	.8508483	.7462994	44.9950625	$\overline{78}.1433916$	1.5664316	36.85	33
6	8.0933	6.5757	.9081256	.8179420	49.3144593	$\overline{78}.2388395$	.9812763	9.58	11
7	9.0933	7.5757	.9587215	.8794228	53.0211921	$\overline{88}.9064037$	.3555733	2.27	2
8	10.0933	8.5757	1.0040332	.9332696	56.2676641	$\overline{86}.0264443$	$\bar{1}.7220859$	.53	1

NORMAL CURVE

The Normal curve has unlimited range and is symmetrical. The equation is

$$Y=\frac{N}{\sigma\sqrt{2\pi}}e^{-x^2/2\sigma^2}$$

For a Normal curve we obtain the μ, β, and κ constants, mean, σ, and Y_0, which equals $\frac{N}{\sigma\sqrt{2\pi}}$.

The following illustration will show the arrangement of columns and the calculation of the values at the ordinates. The data are the result of tossing 8 pennies 2000 times and recording the number of heads obtained. While these data are discontinuous (obtained by counting) they may be used to illustrate a Normal curve.

Number of Heads	f
0	11
1	62
2	196
3	421
4	574
5	467
6	203
7	55
8	11
Total	2000

We obtain

$$\mu_2 = 1.94412$$
$$\mu_3 = -.15724$$
$$\mu_4 = 11.19834$$
$$\beta_1 = .00336$$
$$\beta_2 = 2.96284$$
$$\kappa_2 = -.02989$$

From these values it is apparent that the distribution is a Normal curve.

Mean = 4.01950

$$\sigma = 1.39432$$

$$Y_0 = \frac{2000}{(1.39432)\,(2.506628)} = 572.23952$$

The ordinates for the Normal curve may be calculated directly from the equation

$$Y = \frac{N}{\sigma\sqrt{2\pi}}\,e^{-x^2/2\sigma^2}$$

This may be written as

$$Y = Y_0\,\frac{1}{e^{x^2/2\sigma^2}} \quad \text{since } Y_0 = \frac{N}{\sigma\sqrt{2\pi}}$$

The value for Y_0 may be calculated and then for the various values of x, the deviation of the ordinate from the mean, the second part of the equation, $\frac{1}{e^{x^2/2\sigma^2}}$, may be obtained. This may be done by logarithms and the following arrangement may be used:

$$\frac{\left(\log \frac{1}{e}\right) x^2}{2\sigma^2}$$

Substituting in this equation for the first ordinate of the problem at hand, we have

$$\frac{(9.5657055-10)\,(4.0195)^2}{2(1.39432)^2} = \bar{2}.1954326$$

The number corresponding to logarithm $\bar{2}.1954326$ is .01568. Then

$$Y = (572.23952)\,(.01568) = 8.97$$

The other ordinates may be determined in the same way, or they may be more conveniently calculated by using the z values from Table II in Pearson's Tables. The equation to the curve

$$Y = \frac{N}{\sigma\sqrt{2\pi}}\,e^{-x^2/2\sigma^2}$$

may be divided into two parts, as $\frac{N}{\sigma}$ and $\frac{1}{\sqrt{2\pi}}\,e^{-x^2/2\sigma^2}$

The z values from Pearson's Table II give the second part of this equation in accordance with the varying values of $\frac{x}{\sigma}$. Multiplying these several values for $\frac{x}{\sigma}$ by $\frac{N}{\sigma}$, or 1434.3910, we have the values at the ordinates. Thus for class $V = 3$ we have

$$x = V - M = 3 - 4.0195 = -1.0195 \text{ and } -1.0195/\sigma = -.73118$$

For $x = .73118$ we find from Table II in Pearson's Tables that $z = .3053633$. Multiplying this z value by 1434.3910 we have $Y = 438.01$.

Columns are formed and the various ordinates calculated as follows:

1	2	3	4	5	6
V	f	$V-M$ OR x	x/σ	z VALUES CORRESPONDING TO x/σ FROM PEARSON'S TABLE II	Y COLUMN 5 TIMES N/σ
0	11	-4.0195	-2.88277	.0062570	8.97
1	62	-3.0195	-2.16557	.0382451	54.86
2	196	-2.0195	-1.44838	.1397594	200.47
3	421	-1.0195	- .73118	.3053633	438.01
4	574	- .0195	- .01399	.3988984	572.18
5	467	.9805	.70321	.3115498	446.88
6	203	1.9805	1.42041	.1454797	208.67
7	55	2.9805	2.13760	.0406168	58.26
8	11	3.9805	2.85480	.0067800	9.73

CURVE TYPE II

A curve of Type II is symmetrical with a limited range. The equation to the curve is

$$Y = Y_0\left(1-\frac{x^2}{a^2}\right)^m$$

To calculate the ordinates the following constants are needed:

$$m = \frac{5\beta_2 - 9}{2(3-\beta_2)}$$

$$a^2 = \frac{2\mu_2\beta_2}{3-\beta_2}$$

$$a = \sqrt{a^2}$$

$$Y_0 = \frac{N\,\Gamma(2m+2)}{a\,2^{2m+1}[\Gamma(m+1)]^2}$$

Limits of curve $= M - a$ and $M + a$

For the value $\left(1-\frac{x^2}{a^2}\right)^m$, it is simpler and more convenient to calculate it as $\left[\left(1+\frac{x}{a}\right)\left(1-\frac{x}{a}\right)\right]^m$.

The following data on average number of spikelets per culm in oats will be used to illustrate the method of calculating a Type II curve.

AVERAGE NUMBER OF SPIKELETS	f
10.00–14.99	8
15.00–19.99	28
20.00–24.99	57
25.00–29.99	100
30.00–34.99	115
35.00–39.99	102
40.00–44.99	61
45.00–49.99	26
50.00–54.99	3
TOTAL	500

For this distribution the following constants are obtained, using Sheppard's corrections:

$$
\begin{aligned}
\mu_2 &= 2.5877\\
\mu_3 &= -.4235\\
\mu_4 &= 17.1827\\
\beta_1 &= .0104\\
\beta_2 &= 2.5660\\
\kappa_2 &= -.0088
\end{aligned}
$$

From the necessary constants the values for m, a, and Y_0 are found as follows:

$$
\begin{aligned}
m &= \frac{12.8300-9}{2(3-2.5660)} = 4.4124\\
a^2 &= \frac{(5.1754)\,(2.5660)}{3-2.5660} = 30.5993\\
a &= \sqrt{30.5993} = 5.5317\\
Y_0 &= \frac{500\,\Gamma(8.8248+2)}{(5.5317)\,(2^{8.8248+1})\,[\Gamma(4.4124+1)]^2}
\end{aligned}
$$

The value for Y_0 may be conveniently obtained by logarithms, that is, the sum of the logarithms of the values in the denominator is subtracted from the sum of the logarithms of the values in the numerator. By logarithms we have

$$\log Y_0 = 2.0642153$$

To calculate the ordinates columns are arranged and the work carried out as follows. The ordinates are calculated by determining the deviations from the mean, as in a curve of Type II the mode and mean coincide. In this particular case there is a very slight numerical difference between the mode and mean.

$$\text{Mean} = 32.3400$$

The class value is V and x is the distance from the mean in calculation units of 5 spikelets. In determining x the difference between the mean and the class value is obtained and divided by the class interval, 5, since the calculations are much simpler when based on a unity difference between classes.

DETAILS OF DETERMINING CALCULATED Y VALUES FOR TYPE II CURVE

1	2	3	4	5	6	7	8	9	10	11	
V	x	x/a	$1+x/a$	$1-x/a$	LOG $(1+x/a)$	LOG $(1-x/a)$	COLUMN 6+ COLUMN 7	COLUMN 8 TIMES m	COLUMN 9+ LOG Y_o= LOG Y	CALCULATED Y	OBSERVED Y
12.5	-3.9680	-.7173	.2827	1 7173	9.4513258	.2348462	9.6861720	$\bar{2}$.6152653	.6794806	4.78	8
17.5	-2.9680	-.5365	.4635	1.5365	9.6660497	.1865326	9.8525823	$\bar{1}$.3495341	1 4137494	25.93	28
22.5	-1.9680	-.3558	.6442	1.3558	9.8090207	.1321956	9.9412163	$\bar{1}$.7406228	1 8048381	63.80	57
27.5	– .9680	-.1750	.8250	1.1750	9.9164539	.0700379	9.9864918	$\bar{1}$.9403964	2.0046117	101.07	100
32.5	.0320	.0058	1.0058	.9942	.0025116	9.9974738	9.9999854	$\bar{1}$.9999356	2.0641509	115.92	115
37.5	1.0320	.1866	1.1866	.8134	.0743043	9.9103042	9.9846085	$\bar{1}$.9320865	1.9963018	99.15	102
42.5	2.0320	.3673	1 3673	.6327	.1358638	9.8011978	9.9370616	$\bar{1}$.7222906	1.7865059	61.17	61
47.5	3.0320	.5481	1.5481	.4519	.1897990	9.6550423	9.8448413	$\bar{1}$.3153778	1.3795931	23.97	26
52.5	4.0320	7289	1.7289	.2711	.2377699	9.4331295	9.6708994	$\bar{2}$.5478765	.6120918	4.09	3

CURVE TYPE III

In a curve of Type III the range is limited in one direction only The equation is

$$Y=Y_o\left(1+\frac{x}{a}\right)^{\gamma a}e^{-\gamma x}$$

The following constants are needed

$$\gamma=\frac{2\mu_2}{\mu_3}$$

$$p=\left(\frac{4}{\beta_1}\right)-1$$

$$a=\frac{p}{\gamma}\ \textit{(distance from mode to limited range end)}$$

$$Y_o=\frac{N}{a}\left\{\frac{p^{p+1}}{e^p\Gamma(p+1)}\right\}$$

The origin is at the mode, which is at a distance from the mean given by $\frac{1}{\gamma}$.

The method of calculating a Type III curve is illustrated with the following data, giving the total yield of cat plants in grams.

TOTAL YIELD OF PLANT IN GRAMS	f
0.00 – 1.99	87
2.00 – 3.99	192
4.00 – 5.99	128
6.00 – 7.99	71
8.00 – 9.99	12
10.00 – 11.99	7
12.00 – 13.99	3
TOTAL	500

The following constants are obtained:

$$\mu_2=1.3134$$

$$\mu_2=1.8033$$

$\mu_4 = 7.0092$

$\beta_1 = .7497$

$\beta_2 = 4.0633$

From the necessary constants the values for γ, p, a, and log Y_o are found as follows:

$$\gamma = \frac{2.6268}{1.3033} = 2.0155$$

$$p = \frac{4}{.7497} - 1 = 4.3355$$

$$a = \frac{4.3355}{2.0155} = 2.1511$$

$$Y_o = \frac{500}{2.1511} \frac{4.3355^{5.3355}}{2.7182818^{4.3355}\Gamma(4.3355+1)}$$

Calculating Y_o by logarithms we have:

log $Y_o = 2.2774059$

Mean $= 4.0480$

Mode $= 4.0480 - (\frac{1}{2.0155}$ *times the class interval*$) = 3.0556$

Columns are formed and the work carried out as follows, x being the deviation of the ordinate from the mode expressed on a unity difference between the classes. For ease in calculation the equation for Y is put in the following form:

$$Y = \frac{Y_o}{a^p}(1+x)^p e^{-(x\gamma)}$$

DETAILS OF DETERMINING CALCULATED Y VALUES FOR TYPE III CURVE

1	2	3	4	5	6	7	8	9	
V	x	$a+x$	LOG $(a+x)$	LOG $(a+x)p$	$x\gamma$	LOG $e^{-(x\gamma)}$	COLUMN 5+ COLUMN 7+ LOG Y_0/a^p= LOG Y	CALCULATED Y	OBSERVED Y
1.0	-1.0278	1.1233	.0504958	.2189245	-2.0715	.8996411	1.9537215	89.89	87
3.0	- .0278	2.1233	.3270114	1.4177579	- .0560	.0243205	2.2772343	189.34	192
5.0	.9722	3.1233	.4946137	2.1443977	1.9595	$\bar{1}$.1489999	2.1285535	134.45	128
7.0	1.9722	4.1233	.6152449	2.6673943	3.9750	$\bar{2}$.2736794	1.7762296	59.74	71
9.0	2.9722	5.1233	.7095498	3.0762532	5.9905	$\bar{3}$.3983588	1.3097679	20.41	12
11.0	3.9722	6.1233	.7869855	3.4119756	8.0060	$\bar{4}$.5230382	.7701697	5.89	7
13.0	4.9722	7.1233	.8526812	3.6967993	10.0215	$\bar{5}$.6477177	.1796729	1.51	3

CURVE TYPE V

In a curve of Type V the range is limited at one end only. The equation is

$$Y = Y_0\, x^{-p} e^{-\gamma/x}$$

The following constants are needed:

$$p = 4 + \frac{8 + 4\sqrt{4 + \beta_1}}{\beta_1}$$

$$\gamma = (p-2)\sqrt{\mu_2\,(p-3)}$$

The sign of γ is the same as that of μ_3.

$$Y_0 = \frac{N\gamma^{p-1}}{\Gamma(p-1)}$$

$$\text{Origin} = \text{Mean} - \frac{\gamma}{p-2}.$$

The origin is at the end of the curve.

The method of calculating a Type V curve is illustrated with the following data, giving the average height of oat plants in centimeters.

AVERAGE HEIGHT OF PLANT IN CENTIMETERS	f
45.00 – 49.99	2
50.00 – 54.99	9
55.00 – 59.99	21
60.00 – 64.99	34
65.00 – 69.99	97
70.00 – 74.99	123
75.00 – 79.99	89
80.00 – 84.99	24
85.00 – 89.99	0
90.00 – 94.99	1
TOTAL	400

Constants are obtained as follows:

$$
\begin{aligned}
\mu_2 &= 1.9319 \\
\mu_3 &= -1.4316 \\
\mu_4 &= 13.1697 \\
\beta_1 &= .2843 \\
\beta_2 &= 3.5287 \\
\kappa_2 &= 1.1171
\end{aligned}
$$

From the necessary constants, values for p, γ, and Y_0 are obtained as follows:

$$p=4+\frac{8+4\sqrt{4+.2843}}{.2843}=61.2620$$

$$\gamma=(61.2620-2)\sqrt{1.9319\ (61.2620-3)}=-628.7283$$

$$Y_0=\frac{(400)\ (-628.7283)^{61.2620-1}}{\Gamma(61.2620-1)}$$

Calculating Y_0 by logarithms, we have

$$\log Y_0=90.6358432$$

$$\text{Mean}=70.8375$$

$$\text{Origin}=70.8375-(\frac{-628.7283}{61.2620-2}\ \textit{times the class interval})=123.8840$$

Columns are formed and the work carried out as on page 283, x being the deviation of the ordinate from the origin expressed on a unity difference between the classes.

Logarithmic Curve. There is another form of curve that is useful for certain cases, for example, if one has a series of values in which the first few items may be curving rather rapidly and the remaining values tend to curve less rapidly. In such cases the type of parabola discussed in Chapter VII does not fit so well, and an equation which contains a factor to correct for such differences in curvature, usually some form of a logarithmic curve, will satisfy the conditions better than a second or third order parabola.

There are different forms of logarithmic curves, and the one presented here is particularly adapted to satisfy the condition mentioned above. The observed values to which the curve will be fitted are given on page 284.

DETAILS OF DETERMINING CALCULATED Y VALUES FOR TYPE V CURVE

1	2	3	4	5	6	7	8	9	
V	x	Log x	$\frac{1}{\text{Log } x}$	$\left(\frac{1}{\text{Log } x}\right)^p$	$\frac{\text{Log } e\gamma}{x}$	$\frac{1}{\frac{\text{Log } e\gamma}{x}}$	Column 5+ Column 7+ Log Y_0= Log Y	Calculated Y	Observed Y
47.5	-15.2768	1.1840324	8.8159676	$\overline{73}.4638071$	17.8737198	$\overline{18}.1262802$	.2259305	1.68	2
52.5	-14.2768	1.1546309	8.8453691	$\overline{71}.2650018$	19.1256614	$\overline{20}.8743386$	.7751836	5.96	9
57.5	-13.2768	1.1230934	8.8769066	$\overline{69}.1970521$	20.5661939	$\overline{21}.4338061$	1.2667014	18.48	21
62.5	-12.2768	1.0890852	8.9109148	$\overline{67}.2804625$	22.2414019	$\overline{23}.7585981$	1.6749038	47.30	34
67.5	-11 2768	1.0521859	8.9478141	$\overline{65}.5409874$	24.2137169	$\overline{25}.7862831$	1.9631137	91.86	97
72.5	-10.2768	1.0118579	8.9881421	$\overline{62}.0115613$	26.5698703	$\overline{27}.4301297$	2.0775342	119.55	123
77.5	- 9.2768	.9673982	9.0326018	$\overline{60}.7352515$	29.4339905	$\overline{30}.5660095$	1.9371042	86.52	89
82.5	- 8.2768	.9178625	9.0821375	$\overline{57}.7699075$	32.9901946	$\overline{33}.0098054$	1.4155561	26.03	24
87.5	- 7.2768	.8619404	9.1380596	$\overline{53}.1958072$	37.5238075	$\overline{38}.4761925$	.3078429	2.03	0
92.5	- 6.2768	.7977383	9.2022617	$\overline{49}.1289563$	43.5019823	$\overline{44}.4980177$	$\overline{2}$.2628172	.02	1

y
5
7
9
10
11
12
13

It is noted that the first three observations tend to rise rather rapidly while there is a lessening of this tendency in the last four observations.

The equation to this logarithmic curve is

$$y = a + bx + c \log x$$

in which x is the distance of any ordinate from the origin, a is a constant term, and b and c together determine the direction of the curve. Since there are three unknowns it will be necessary to have three equations, as follows:

EQUATION I $\Sigma a + \Sigma(x)b + \Sigma(\log x)c = \Sigma y$
EQUATION II $\Sigma(x)a + \Sigma(x^2)b + \Sigma(x \log x)c = \Sigma(xy)$
EQUATION III $\Sigma(\log x)a + \Sigma(x \log x)b + \Sigma(\log x)^2 c = \Sigma(y \log x)$

For convenience in making the calculations, columns may be arranged as follows:

$$y \quad a \quad x \quad xy \quad x^2 \quad (\log x) \quad (x \log x) \quad (\log x)^2 \quad (y \log x)$$

The values for columns $(\log x)$, $(x \log x)$, and $(\log x)^2$ may be obtained from Table II in the Appendix, and the values in the last column are obtained by multiplying $\log x$ by the corresponding y value. These various columns are summed and the values substituted in the equations. The steps are as follows, reading to seven decimals:

	y	a	x	xy	x^2	$(\log x)$	$(x \log x)$	$(\log x)^2$	$(y \log x)$
	5	a	1	5	1	.0000000	.0000000	.0000000	.0000000
	7	a	2	14	4	.3010300	.6020600	.0906191	2.1072100
	9	a	3	27	9	.4771213	1.4313639	.2276447	4.2940917
	10	a	4	40	16	.6020600	2.4082400	.3624762	6.0206000
	11	a	5	55	25	.6989700	3.4948500	.4885591	7.6886700
	12	a	6	72	36	.7781513	4.6689078	.6055194	9.3378156
	13	a	7	91	49	.8450980	5.9156860	.7141906	10.9862740
TOTAL	67	7	28	304	140	3.7024306	18.5211077	2.4890091	40.4346613

Substituting these values in the equations, we have

Equation I $7a+28b+3.7024306c=67$
Equation II $28a+140b+18.5211077c=304$
Equation III $3.7024306a+18.5211077b+2.4890091c=40.4346613$

Solving for c

Equation II $28a+140b+18.5211077c=304$
Equation I ×4 $\underline{28a+112b+14.8097224c=268}$
Subtracting $28b+3.7113853c=36$ (1)

III×7 $25.9170142a+129.6477539b+17.4230637c=283.0426291$
I×3.7024306 $\underline{25.9170142a+103.6680568b+13.7079923c=248.0628502}$
Subtracting $25.9796971b+3.7150714c=34.9797789$ (2)

(2)×28 $727.4315188b+104.0219992c=979.4338092$
(1)×25.9796971 $\underline{727.4315188b+96.4206659c=935.2690956}$
Subtracting $7.6013333c=44.1647136$
$c=5.81$

Substituting in (1)

$$28b+21.5631=36$$
$$28b=14.4369$$
$$b=.5156$$

Substituting b and c in I

$$7a+14.4368+21.5111=67$$
$$7a=31.0521$$
$$a=4.4360$$

The equation to the curve is therefore

$$y=4.4360+.5156x+5.81\log x$$

Substituting the various values of x and $\log x$ we have

		CALCULATED y	OBSERVED y
For x_1	4.4360+ .5156+5.81 (.0000000)=	4.95	5
x_2	4.4360+1.0312+5.81 (.3010300)=	7.22	7
x_3	4.4360+1.5468+5.81 (.4771213)=	8.75	9
x_4	4.4360+2.0624+5.81 (.6020600)=	10.00	10
x_5	4.4360+2.5780+5.81 (.6989700)=	11.07	11
x_6	4.4360+3.0936+5.81 (.7781513)=	12.05	12
x_7	4.4360+3.6092+5.81 (.8450980)=	12.96	13

This type of logarithmic curve may often fit the observed values better than a simple parabola and may be found especially useful for certain types of data, such as the application of fertilizers to crops or with some growth curves. In connection with the logarithmic curve, as well as with other curves, one may often find it necessary to fit more than one type and then select the one that seems to give the best results or, as we may term it, the best graduation for the several observed values.

CHAPTER XI

GOODNESS OF FIT

In Chapter X various methods have been given for fitting frequency curves to different types of frequency distributions. Some of the observed values agreed very well with the calculated values, and in other cases there was not such close agreement. Pearson has shown that it is often important in studies of this sort to determine how well the observed frequency agrees with the theoretical curve or frequency. This has been termed the Goodness of Fit, or Chi-square test, and is based on the comparison of the entire theoretical distribution with the observed distribution. The Goodness of Fit or Chi-square test may be applied to other comparisons and it is the purpose of this chapter to give only a few illustrations of the application.

Elderton has prepared a table which enables us to determine the probability, P, that measures the agreement, or Goodness of Fit, between the observed results and the calculated results. For example, suppose we have results that lead to a P value of .75. This indicates a good agreement between the observed and calculated results, since in cases of perfect agreement the value of P cannot be more than 1.00. In order to make the comparison and obtain the value for P we proceed as follows. The differences between the observed frequencies and the calculated frequencies are determined, and a value Chi square, designated χ^2, is obtained from the following relation:

$$\chi^2 = \Sigma \frac{(o-c)^2}{c}$$

In this formula o refers to the observed frequencies and c to the calculated frequencies. The difference between the observed and calculated frequency for each class is obtained, squared, and divided by the calculated frequency, and the sum of these quotients is the value for χ^2.

An application of this test for Goodness of Fit will be made with the calculated values for the Type I curve as given in Chapter X. The various steps are given in Table 76.

TABLE 76

APPLICATION OF GOODNESS OF FIT TO RESULTS OBTAINED FROM FITTING A TYPE I CURVE

OBSERVED o	CALCULATED c	$(o-c)$	$(o-c)^2$	$\frac{(o-c)^2}{c}$
3	4.34	−1.34	1.7956	.414
50	46.57	3.43	11.7649	.253
106	105.54	.46	.2116	.002
109	115.52	−6.52	42.5104	.368
80	78.19	1.81	3.2761	.042
42	35.78	6.22	38.6884	1.081
7	11.29	−4.29	18.4041	1.630
2	2.39	− .39	.1521	.064
1	.32	.68	.4624	1.445
				$\chi^2=5.299$

From Elderton's table

For $n'=9$ and $\chi^2=5$ $P=.757576$

For $n'=9$ and $\chi^2=6$ $P=.647232$

Difference .110344

Interpolating between $\chi^2=5.299$ and $\chi^2=5$ we have

$.757576-(.110344\times .299)=.724583$

Therefore $P=.725$

The observed values appear in the first column of the table and the calculated values obtained for the same classes are given in the second column. The differences between these observed and calculated values, or *o-c*, are recorded in the third column, and the squares of these differences are given in the fourth column. These squared differences are divided by the calculated or expected values for the corresponding classes, giving the results in the last column. Summing these values we find $\chi^2=5.299$.

In order to interpret this result one should have access to Elderton's table, which is included in *Pearson's Tables for Statisticians and Biometricians.* In this table of P values the values of X^2 are given at the left of the table for the values of n' as indicated at the head of the columns. Here n' refers to the number of classes or groups. A table for testing Goodness of Fit is also included in *Fisher's Statistical Methods for Research Workers.* This table is arranged in a different manner than Elderton's table in that the rows are indicated by n, which refers to the degrees of freedom, and the numbers at the head of the columns give the probability values, P, while the values in the table are those for X^2.

For the problem at hand we find from Elderton's table for 9 groups and a X^2 value of 5.299 that the probability is .725, which indicates a very good agreement or fit. That is, due to chance variation or random sampling we may expect such a deviation from theory 72 out of 100 times.

The question naturally arises as to what value of P may be accepted as indicating a good fit or, in other words, what may be considered the division between a good fit and a poor fit. Elderton says: "It is impossible to fix such a value. We have merely a measure of probability for the whole table, and if the odds against the graduation are twenty or thirty to one the result is unsatisfactory; if they are ten to one the graduation is not unreasonable, but the exact value when a result must be discarded cannot be given." Fisher states: ". . . we do not want to know the exact value of P for any observed X^2, but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05, and consider that higher values of X^2 indicate a real discrepancy." What we have is a measure of the comparison between the observed and calculated values for the whole distribution, and there are times when the value of P may be low and yet the calculated curve in general may fit the data very well. That is, the value of P may be low due to the comparatively high value of X^2, which may be rendered high by unusual deviations which sometimes occur in the more extreme classes. For this reason it may often be better to group together the extreme classes of a frequency distribution, since it is desirable that if possible the number in any class should not be less than 10. When the data in Table 76 are grouped so that no class has less than 10 individuals, the first two classes would be combined, giving 53 individuals, and the last three classes would be combined, giving 10 individuals. The resulting X^2 would then be 3.494 and P would be .626.

The method just described is the one that was used originally and is still used by many investigators when making application of the Chi-square test to the comparison of frequencies. It has been pointed out by Fisher that it is better to take n not as the number of groups but as the number of degrees of freedom, which is the number of groups minus the number of constants that have been used in the fitting. For example, if a normal curve were fitted to a frequency distribution having 10 classes, where the mean and standard deviation have been determined in fitting the curve and where the totals of the two distributions have been made equal or approximately so, n would equal the number of groups minus 1 degree of freedom for each of these constants. In this case, n would be 10–3. In using Elderton's table to obtain the P value we would seek the value of P for $n' = n + 1$. In other comparisons between two distributions which are made to agree in totals only, the degrees of freedom will equal the number of classes or groups minus 1. In such comparisons as Mendelian studies where the material is grouped in say four classes, it is assumed that three of these classes may be filled in arbitrarily and therefore the degrees of freedom will equal 3.

It is desirable to have access to Fisher's or Elderton's tables in order to evaluate properly the results obtained by the Chi-square test. When these tables are not available the following values will serve as a guide in interpretation.

VALUES FOR INTERPRETING GOODNESS OF FIT

1 DEGREES OF FREEDOM n	2 χ^2 VALUE FOR HIGH PROBABILITY	3 χ^2 VALUE FOR FAIR PROBABILITY
1	.2	3.0
2	.7	4.7
3	1.4	6.3
4	2.2	7.8
5	3.0	9.3
6	3.8	10.7
7	4.7	12.0
8	5.5	13.4
9	6.4	14.7
10	7.3	16.0
11	8.1	17.3
12	9.0	18.6
13	9.9	19.8
14	10.8	21.1
15	11.7	22.3
16	12.6	23.6
17	13.5	24.8
18	14.4	26.0
19	15.3	27.2
20	16.3	28.4
21	17.2	29.6
22	18.1	30.9
23	19.0	32.0
24	19.9	33.1
25	20.9	34.5
26	21.8	35.5
27	22.7	36.7
28	23.6	38.0
29	24.6	39.0
30	25.6	40.4

Two sets of values are given for χ^2 for each degree of freedom, n, and they may be interpreted in the following way. When a Chi-square determination leads to a value equal to or less than the value of χ^2 in the second column, for the degrees of freedom concerned, the fit is very good, or there is a close agreement between the results compared. When the Chi-square determination is more than that tabled in the second column, but equal to or less than that in the

TABLE 77

APPLICATION OF GOODNESS OF FIT IN A COMPARISON OF FREQUENCY DISTRIBUTIONS*

1	2	3	4	5	6	7	8	9
CLASS	f_A	f_B	f_A+f_B	f_A/N_A	f_B/N_B	$f_A/N_A-f_B/N_B$	$(f_A/N_A-f_B/N_B)^2$	$\frac{(f_A/N_A-f_B/N_B)^2}{f_A+f_B}$
9	3	1	4	.0476	.0143	.0333	.00110889	.0002772
10	8	5	13	.1270	.0714	.0556	.00309136	.0002378
11	14	8	22	.2222	.1143	.1079	.01164241	.0005292
12	13	9	22	.2063	.1286	.0777	.00603729	.0002744
13	16	25	41	.2540	.3571	-.1031	.01062961	.0002593
14	8	20	28	.1270	.2857	-.1587	.02518569	.0008995
15	1	2	3	.0159	.0286	-.0127	.00016129	.0000538
	$N_A=63$	$N_B=70$						$\Sigma=.0025312$

$\chi^2=.0025312\times 63\times 70=11.1626$ $P=.084082$

---

*Data from R. A. Emerson

third column, there is a fair agreement. In other words, the values of χ^2 in the second column lead to a high value of P while those in the third column lead to a fair value of P. For Chi-square determinations higher than those in the third column there is not a close agreement between the comparisons, or between observation and theory.

When n is more than 30 the results may be compared by assuming that $\sqrt{2\chi^2}$ is distributed normally about a mean equal to $\sqrt{2n-1}$ with unit standard deviation. For example, if $\chi^2=60.5$ and $n=41$ we have

$$\begin{aligned} \sqrt{2\chi^2} &= \sqrt{2(60.5)} = 11.00 \\ \sqrt{2n-1} &= \sqrt{2(41)-1} = 9.00 \\ &\text{Difference} = 2.00 \pm 1.00 \end{aligned}$$

The difference is twice the probable error and borders on significance.

The application of Goodness of Fit is very useful in problems of curve fitting when there may have been a question as to which particular type of curve should have been used for the graduation of the data. If one is in doubt it is well to calculate the expected frequencies from the types of curves that seem to fit the observed frequencies and by determining the Goodness of Fit select the type that gives the highest value for P. When much curve fitting is being done there are times when it is preferable, for simplicity in calculation, to use a curve that may give a slightly poorer fit, or a lower value for P, but in general the best fitting type should be used.

The Goodness of Fit idea may be extended further to compare one frequency distribution with another to determine whether they may have arisen as samples from the same population, or, in other words, this method may be used when applied to frequency distributions to determine the homogeneity of the material at hand. The method is very useful in connection with certain problems in inheritance when one may be investigating quantitative characters and the results are being compared by means of frequency distributions. The value of χ^2 may be determined from the formula

$$\chi^2=\left[\Sigma\,\frac{(f_A/N_A-f_B/N_B)^2}{f_A+f_B}\right]N_A\,N_B$$

in which f_A and f_B are the frequency distributions and N_A and N_B are the totals of these frequencies. The various calculations are simple, and a convenient arrangement for determining χ^2 is illustrated in Table 77, page 292.

The problem here is a study of the inheritance of length of ears in corn. The first column of the table gives the class length in centimeters, and the second and third columns give the distributions for two third-generation families. The problem is to determine whether these two frequency distributions are really different. If they are different, we may conclude that they have a different genetic constitution. If they are apparently the same, then we may conclude that they may be the same in their genetic constitution.

The various steps are as follows. In column 4 are recorded the sums of the two frequency distributions and in columns 5 and 6 the frequencies of the individual distributions are expressed as a percentage of the total. In columns 7 and 8 the differences pair by pair have been obtained and squared, and these squared values are divided by the corresponding values in column 4 and recorded in column 9. The values in column 9 are summed, and χ^2 is obtained from this summation value multiplied by N_A N_B, the totals of the frequencies. The value for χ^2 determined by this method is 11.1626, and there are 7 groups, giving 6 degrees of freedom.

This may be interpreted in the manner already explained. Referring to Elderton's table we will have for $n' = 6+1$, or 7

For $n' = 7$ and $\chi^2 = 11$	$P = .088376$
For $n' = 7$ and $\chi^2 = 12$	$P = .061969$
Difference	.026407

Our value for χ^2 is 11.1626, and interpolating between this and the P value for $\chi^2 = 11$, we have $.088376 - (.026407 \times .1626)$, or a P value of .084082. This gives a rather low value for the probability, and yet not low enough to be absolutely certain that the two frequency distributions are really samples from different populations. Using this value of P to determine odds we find that the odds are about 11 to 1, which are not quite high enough to indicate significance.

If, on the other hand, we interpret the results for these data on the basis of the means and their probable errors alone, without considering the entire distributions, we have

Mean of series B	12.71±.11
Mean of series A	11.94±.12
Difference	.77±.16

The difference is nearly five times its probable error, and we would conclude that the means are really different. However, it must be pointed out that for comparisons of this sort we are not interested in the mean alone but in the result of the distribution as a whole. In some cases the means may differ and yet a comparison of both distributions as a whole may not indicate a significant difference. Again, the means may be practically the same and still the distributions as a whole may not agree. The real purpose of the Chi-square method is to compare the entire distributions, and when the probability is very low we may conclude that the distributions are really samples from different populations, or when studying genetic problems we would conclude that the two distributions differ in their genetic constitution. When P is high we may conclude that the samples may have arisen from the same population, or that the genetic constitution of the two populations is the same.

Another application of the Goodness of Fit method is in the study of problems of Mendelian segregation, since it is possible to compare the actual observed numbers obtained for the different groups with the expected numbers on the basis of the hypothesis that is used for the interpretation of the results. Such an application is made with the following data obtained in a study of the inheritance of color in oats.

COLOR OF GRAIN	OBSERVED o	CALCULATED c	$(o-c)$	$(o-c)^2$	$\frac{(o-c)^2}{c}$
Black	230	231	–1	1	.0043
Gray	57	58	–1	1	.0172
Yellow	21	19	2	4	.2105
					$\chi^2 = .2320$

The observed values are given, together with the calculated values for the indicated 12: 3: 1 ratio. The steps for determining the Goodness of Fit are carried out as before, and the value for X^2 is found to be .2320. Values for X^2 less than 1 are not included in

Elderton's table and we may therefore refer to Fisher's table. In this example we have three groups or classes and it is assumed that the frequencies in two of these may be distributed arbitrarily, and thus the degrees of freedom are 2. In making general interpretations, Fisher states that n should refer to the degrees of freedom rather than to the number of groups. For example, in comparing the two frequency distributions for length of ears in corn, following Fisher we would read n as 6.

In the problem at hand for $n=2$ we find from Fisher's table that when $P=.90$ the value of χ^2 is .211, and when $P=.80$ the value of χ^2 is .446. The value obtained, .2320, lies between these values but nearer the value for $P=.90$. The probability therefore lies between .90 and .80. This is a high probability and indicates a good agreement between theory and observation.

These examples illustrate how the Goodness of Fit may be applied to results obtained from curve fitting or from similar comparisons. We should be careful in the interpretation of results on the basis of Goodness of Fit. As already indicated, it is not possible to name a definite value above which a good fit exists and below which the fit is poor, since we are dealing with frequency distributions or similar material and the divergencies of a single group may be such as to have an undue influence on χ^2 and therefore on the value of P. When these facts are kept in mind then one may make application of the Goodness of Fit, and should be able to give a satisfactory interpretation to his data.

In presenting these illustrations it has been the purpose to show some of the applications of Goodness of Fit. For a more complete discussion the reader is referred to the most recent edition of Fisher's book, *Statistical Methods for Research Workers.* Since writing the present chapter a further contribution to the problem of Goodness of Fit has been made by Pearson in his paper in *Biometrika,* Volume 26.

CHAPTER XII

ANALYSIS OF SMALL SAMPLES AND APPLICATION OF PROBABILITY

The first part of this book has dealt with the analysis of frequency curves, correlation, and the fitting of curves for large populations. There is a very important phase of statistical analysis in which the agronomist and biologist are interested, and that is the methods that may be used for evaluating results obtained from small samples, since as a rule these investigators are limited as to the number of experiments they can conduct. During recent years this phase of statistical analysis has been receiving considerable attention. It is the purpose to set forth in this and following chapters the methods that may be used, with illustrations of their application and limitation. The discussions will deal with the application of the various methods rather than strictly mathematical considerations. It should be understood that while the examples will be drawn largely from agronomic material, the methods are in no sense limited to this field but may be used for the study of other problems, such as the results obtained in nutrition experiments, the effect of various treatments for insect and disease control, and other problems of similar nature.

In Chapter IX we discussed the meaning of probable error, and gave formulas for determining probable errors for various constants that are used in biometrical work. For the calculation of probable errors for small samples, or for the analysis of the results of small samples, a number of formulas or methods for analysis have been developed. These will be discussed in more or less chronological order.

Bessel's and Peters' Formulas. The first methods to be generally used for the determination of probable errors for small samples are the applications of the formulas developed by Bessel and Peters. It will be recalled that the probable error of a single determination is obtained from $\pm .6745\sigma$. When it is remembered that the formula for the standard deviation of a frequency distribution when working from the true mean is

$$\sigma=\sqrt{\frac{\Sigma fD^2}{N}}$$

It must be clear that the formula for the probable error of a single determination is

$$\pm .6745\sqrt{\frac{\Sigma fD^2}{N}}$$

When working with ungrouped material this formula for the probable error of a single determination is

$$\pm .6745\sqrt{\frac{\Sigma D^2}{N}}$$

It was pointed out in Chapter IX that in our experimental work it is not possible to know the true value, and that the mean or other constant as determined, with its probable error, gives the most probable value. We do not know the true value of a mean, since we do not know the value of all the true errors. As the number in the population increases so that N becomes very large, then the observed errors approach the true errors, or the sum of the squares of the observed errors approaches the sum of the squares of the true errors. When N is small the sum of the squares of the known errors does not approach so closely the sum of the squares of the true errors, and it is a fact that the sum of the squares of the known errors for small samples is less than the sum of the squares of the true errors.

Various corrections for the observed errors have been suggested, but it has been found by actual trial and proved mathematically that if $\frac{\Sigma D^2}{N}$ represents the average square of the true errors for large samples, $\frac{\Sigma D^2}{N-1}$ will approach very closely the average square of the true errors for small samples. This approximation is used in the formula suggested by Bessel for obtaining the probable error of a single determination, which is

$$P.E._s = \pm .6745\sqrt{\frac{\Sigma D^2}{N-1}}$$

It will be noted that the only difference between this formula for the probable error of a single determination and the formula suggested in Chapter IX is the correction for small numbers made by using $N-1$ in the denominator rather than N. In Chapter IX it was shown that the probable error of the mean of a number of determinations may be obtained by dividing the probable error of a single determination by the square root of the number of observations, so for small samples the probable error of the mean may be obtained by dividing the probable error of a single determination

TABLE 78

METHOD OF CALCULATING PROBABLE ERRORS BY BESSEL'S AND PETERS' FORMULAS

YIELD	DEVIATION FROM MEAN D	D^2
38	0	0
40	2	4
40	2	4
42	4	16
39	1	1
35	− 3	9
32	− 6	36
28	−10	100
42	4	16
44	6	36
380	$\Sigma+D=38$	$\Sigma D^2=222$

Mean = 38

By Bessel's formulas

$$P.E._s=\pm .6745\sqrt{\frac{\Sigma D^2}{N-1}}=\pm .6745\sqrt{\frac{222}{9}}=\pm 3.35$$

$$P.E._M=\pm .6745\sqrt{\frac{\Sigma D^2}{N(N-1)}}=\pm .6745\sqrt{\frac{222}{90}}=\pm 1.06$$

By Peters' formulas

$$P.E._s=\pm .8453\frac{\Sigma+D}{\sqrt{N(N-1)}}=\pm .8453\frac{38}{\sqrt{90}}=\pm 3.39$$

$$P.E._M=\pm .8453\frac{\Sigma+D}{N\sqrt{N-1}}=\pm .8453\frac{38}{10\sqrt{9}}=\pm 1.07$$

by the square root of N, the correction for small numbers having been made in the formula for the probable error of a single determination.

It is also possible to obtain the probable error of the mean directly from the formula

$$P.E._M = \pm .6745\sqrt{\frac{\Sigma D^2}{N\,(N-1)}}$$

which is Bessel's formula for the probable error of the mean. These formulas may now be applied to the data in Table 78, page 299, which are the yields of ten plots sown to the same variety of wheat.

The first step is to obtain the mean by the usual process of summing the individual items and dividing by the number of items The deviation, D, of each item from the mean is determined, as given in the second column of the table. These values are squared and the sum obtained. In this example the mean is 38 and the sum of the squares is 222. This sum may now be substituted in the formula for the probable error of a single determination

$$P.E._s = \pm .6745\sqrt{\frac{\Sigma D^2}{N-1}}$$

Since $N = 10$, we substitute 9 in the formula and complete the operations, obtaining the value ±3.35. This is the probable error of a single determination, which in this case is a single yield.

It will be recalled that in the discussion of the probable error in Chapter IX it was pointed out that one-half of the observations should fall within the limits of the value of the probable error of a single determination when this value is subtracted from or added to the mean. This will not hold exactly for small samples, but it will be noted on comparing the values of D in the second column of Table 78 with the probable error of a single determination that one-half of the deviations, namely, the first, second, third, fifth, and sixth, come within the limits of ±3.35.

By substituting the necessary values in Bessel's formula for the probable error of the mean, we find the probable error of the mean is ±1.06. As stated above, this value may also be obtained by the relation $\frac{P.E._s}{\sqrt{N}}$, and $3.35/\sqrt{10} = 1.06$.

It is apparent that the only difference between this method for obtaining the probable error and that discussed in Chapter IX is that a correction factor has been substituted for small numbers.

The sum of the squares is divided by $N-1$ rather than by N, and this results in a larger value from which to extract the square root, giving a slightly larger error.

It may be convenient at times to calculate the errors by Bessel's formulas directly from the values of the individual items without determining the second column of Table 78. From the column of yields the mean is obtained as before, but instead of taking the deviations of the items from the mean the items themselves are squared and summed. The mean is squared and multiplied by N and this value is subtracted from the sum of the squares of the

TABLE 79

METHOD OF CALCULATING PROBABLE ERRORS DIRECTLY FROM THE INDIVIDUAL ITEMS

YIELD OR V	V^2
38	1444
40	1600
40	1600
42	1764
39	1521
35	1225
32	1024
28	784
42	1764
44	1936
TOTAL 380	$\Sigma V^2 = 14662$

Mean = 38

$M^2N = (38)^2 \times 10 = 14440$

$$P.E._s = \pm .6745\sqrt{\frac{\Sigma V^2 - M^2N}{N-1}}$$

$$P.E._s = \pm .6745\sqrt{\frac{14662 - 14440}{9}} = \pm 3.35$$

$$P.E._M = \pm .6745\sqrt{\frac{\Sigma V^2 - M^2N}{N(N-1)}}$$

$$P.E._M = \pm .6745\sqrt{\frac{14662 - 14440}{90}} = \pm 1.06$$

several observations. The difference is divided by $N-1$ and the square root extracted. This is multiplied by $\pm .6745$, giving the probable error of a single determination. The formula may be written as follows:

$$P.E._s = \pm .6745\sqrt{\frac{\Sigma V^2 - M^2 N}{N-1}}$$

in which ΣV^2 is the sum of the squares of the several items, or values, and M is the mean. The details of the method are shown in Table 79, page 301, with the data from Table 78. The result is

$$P.E._s = \pm .6745\sqrt{\frac{14662-14440}{9}} = \pm .6745\sqrt{24.666667} = \pm 3.35$$

This gives the same value under the radical and the final result is the same as obtained by the other method.

The probable error of the mean may also be obtained by squaring the items directly and using the following formula:

$$P.E._M = \pm .6745\sqrt{\frac{\Sigma V^2 - M^2 N}{N\,(N-1)}}$$

Substituting the necessary values from Table 79 we have

$$P.E._M = \pm .6745\sqrt{\frac{14632-14440}{90}} = \pm 1.06$$

It may also be pointed out that the probable error of a single determination may be obtained by adapting one of the methods suggested in Chapter V for determining the standard deviation. This formula for the probable error of a single determination is

$$P.E._s = \pm .6745\sqrt{\frac{N\Sigma V^2 - (\Sigma V)^2}{N\,(N-1)}}$$

Using the data in Table 79 we have for the summation of V^2, 14662, and for $(\Sigma V)^2$, 144400, and N is 10. Substituting these values in the formula we have

$$P.E._s = \pm .6745\sqrt{\frac{(10)\ (14662)\ -144400}{(10)\qquad (9)}} = \pm 3.35$$

The probable error of the mean may be obtained from $3.35/\sqrt{N}$, or $P.E._M = \pm 1.06$.

One advantage of these methods is that they make for greater accuracy in not dropping decimals. For example, with the usual method when the mean is not obtained exactly and decimals are dropped, or a number in the deviation column is raised on account of the value of certain decimals, slight errors may occur. An objection to these methods is that when the values observed are large it is necessary to deal with large numbers in the calculations, but with the usual helps in statistical work this is not a serious matter.

The formulas that Peters has suggested are based on the deviations themselves and not on the squares of the deviations, as in Bessel's formulas. Peters' formula for the probable error of a single determination is

$$P.E._s = \pm .8453 \frac{\Sigma + D}{\sqrt{N(N-1)}}$$

in which $+D$ means that all the deviations from the mean are summed without regard to signs. This formula may be applied to the data in Table 78. The mean and the deviations from the mean are obtained, and the sum of these deviations, $\Sigma + D$, is 38. Substituting the necessary values in the formula, we have

$$P.E._s = \pm .8453 \frac{38}{\sqrt{90}} = \pm 3.39$$

Peters' formula for the probable error of the mean is

$$P.E._M = \pm .8453 \frac{\Sigma + D}{N\sqrt{N-1}}$$

Again substituting the values from Table 78 we have

$$P.E._M = \pm .8453 \frac{38}{10\sqrt{9}} = \pm 1.07$$

The constant used in these formulas is ±.8453 while in Bessel's formulas it is ±.6745. This is due to the fact that there is a relation between these two formulas similar to that between the standard and average deviations. In the discussion on constants of dispersion it was stated that, for symmetrical distributions, the average deviation is equal to .7979 *S.D.*, and from this we derive the value .8453 from the following relation

$$A.D. : S.D. :: .7979 : 1$$

Then from

$$.7979 : 1 :: .6745 : x$$

we have

$$x = .8453$$

Therefore, when using Peters' formula, it is necessary to multiply the value obtained from $\frac{\Sigma + D}{\sqrt{N(N-1)}}$ by .8453 in order to include one-half of the individuals, or, more generally, one-half of the area of the curve.

In general, the results obtained from the two methods will agree very closely, as is true in the example just given. A comparison of the results obtained with the two sets of equations and using the same data is given in Table 80.

Table 80

Comparison of Probable Errors of a Single Determination and of the Mean When Calculated by Bessel's and Peters' Formulas

Probable Error of a Single Determination		Probable Error of the Mean	
Bessel's Formula	Peters' Formula	Bessel's Formula	Peters' Formula
4.12	4.32	1.30	1.37
4.82	4.71	1.52	1.48
3.78	3.72	1.20	1.18
2.48	2.44	.78	.77
4.84	5.02	1.53	1.59
3.28	3.62	1.04	1.14
3.91	3.78	1.24	1.20
4.46	4.81	1.41	1.52
2.62	2.98	.83	.94
3.73	3.89	1.18	1.23
Ave. 3.80	3.93	1.20	1.24

From these values it is apparent that the results obtained from the two methods agree very closely in most cases. For the results in Table 80 the ratio of those from Bessel's formulas compared with those from Peters' formulas is about 1.00 to 1.03. For a larger number of comparisons the ratio has been found to be about 1.00 to 1.07.

Of the two methods, Bessel's formulas insure greater accuracy. To facilitate the calculation of probable errors by Bessel's and Peters' formulas, Tables IV and V are included in the Appendix for the values $\frac{.6745}{\sqrt{(N-1)}}$, $\frac{.6745}{\sqrt{N(N-1)}}$, $\frac{.8453}{\sqrt{N(N-1)}}$, and $\frac{.8453}{N\sqrt{N-1}}$ for all numbers of N from 1 to 99. Obtaining from these tables the value corresponding to the number of observations, it is only necessary to multiply this value by the square root of ΣD^2 to obtain the required probable error by Bessel's formulas, and by $\Sigma + D$ when using Peters' formulas.

The interpretation of results on the basis of probable errors calculated in this way, and the calculation of odds, are the same as already explained.

Comparison of Differences. We may now proceed to compare two results on the basis of the probable errors of the means obtained by Bessel's formula, using the data in Table 81, page 306. These data are the yields from ten plots each of two varieties of wheat.

The data for variety A are the same as given in Table 78 and the mean and probable error of the mean have already been determined. By the same process the mean and its probable error for variety B are found to be 34.00±1.16. Applying the formula for the probable error of a difference as given in Chapter IX,

$$P.E._{\text{Difference}} = \sqrt{E_1^2 + E_2^2} \text{ or } \sqrt{a^2 + b^2}$$

we have

Variety A	38.00±1.06
Variety B	34.00±1.16
Difference	4.00±1.57

To determine whether or not this difference is significant we divide the difference by its probable error and determine the odds.

TABLE 81

COMPARISON OF TWO MEANS ON THE BASIS OF THEIR PROBABLE ERRORS

VARIETY A	VARIETY B
38	37
40	37
40	40
42	40
39	32
35	30
32	31
28	22
42	36
44	35
MEAN = 38.00	MEAN = 34.00

Variety A 38.00±1.06
Variety B 34.00±1.16

Difference 4.00±1.57

$$\frac{D}{P.E.} = \frac{4.00}{1.57} = 2.55$$

Odds 10.70 to 1

The difference divided by its probable error, or 4.00 /1.57, is 2.55. As explained in Chapter IX, this $\frac{D}{P.E.}$ ratio is multiplied by .6745, giving 1.720, and from Table VI in the Appendix we find the odds to be 10.70 to 1. These odds may also be read directly from Table VII in the Appendix. In Chapter IX it was stated that odds of 30 to 1 are usually accepted as indicating significance. We would therefore conclude that varieties A and B are not significantly different since the odds are only 10.70 to 1.

In following the standard of accepting odds of 30 to 1 as indicating significance it should be understood that it is not possible to say absolutely that odds of 30 to 1 are significant, while a result leading to odds of 29 to 1 should be discarded as being of no significance. The use of odds of 30 to 1 has been adopted as a convenient division point between significance and non-significance, but extreme caution should be observed in drawing conclusions from comparisons when the odds are only 30 to 1. In the calculations the method of handling decimals or carrying the calculations to more decimals may at times change the odds from 29 to 1 to 30 to 1.

Considering further the difference between A and B, it may be well to call attention to the fact that the formula used for the probable error of a difference is only part of the complete formula, as will now be explained. Suppose we have a series of measurements, or a series of yields, for A and B. Let D_A represent the deviations of the individual items of A from the mean of A, and let D_B represent the deviations of the individual items of B from the mean of B. Suppose that the differences between the yields of the two series are obtained, and from these the average difference is determined. Then let $D_{A\text{-}B}$ be the deviations of the differences from the average difference. Then

$$D_{A\text{-}B} = D_A - D_B$$

Squaring

$$D^2_{A\text{-}B} = D^2_A + D^2_B - 2D_A D_B$$

Summing all these deviations we have

$$\Sigma D^2_{A\text{-}B} = \Sigma D^2_A + \Sigma D^2_B - \Sigma 2D_A D_B$$

Dividing by N

$$\frac{\Sigma D^2_{A\text{-}B}}{N} = \frac{\Sigma D^2_A}{N} + \frac{\Sigma D^2_B}{N} - \frac{\Sigma 2D_A D_B}{N}$$

Now

$\frac{\Sigma D^2_{A\text{-}B}}{N}$, $\frac{\Sigma D^2_A}{N}$, and $\frac{\Sigma D^2_B}{N}$ are the squares of the standard deviations, and

$\frac{\Sigma 2D_A D_B}{N}$ is twice the mean product deviation, or $\Sigma P/N$ of the formula for the correlation coefficient

$$r = \frac{\Sigma D_A D_B}{N\sigma_A \sigma_B}, \text{ or } r\sigma_A\sigma_B = \frac{\Sigma D_A D_B}{N}$$

Substituting in the formula above we have

$$\sigma^2_{A\text{-}B} = \sigma^2_A + \sigma^2_B - 2r\,(\sigma_A\sigma_B)$$

This is the complete formula to be used when comparing a series of results, as for example a series of measurements of A and a series of measurements of B. If there is correlation between the results it is necessary to eliminate the effect of correlation in order to have the true value for the probable error. It is evident from the formula that when a correlation is positive the probable error as obtained will be smaller than if the correlation were negative. In other words, when the correlation is positive we are subtracting a positive value, and when the correlation is negative we are subtracting a negative value, which actually means that we add in the effect of correlation, and hence the standard deviation of the difference, and the probable error, will be larger.

Correlation between results of experiments may be due to various factors which operate to produce similar changes in the variables studied. In the example before us we are dealing with the yields of two varieties of wheat when each variety is grown in ten replicated plots. The field on which these varieties were grown varies so far as its fertility is concerned, and so it is evident that when a plot of A and a plot of B are grown in one part of the field the yields may be relatively high or low depending on the fertility. We see by inspection that there is a tendency for the yields to rise and fall together. How much correlation is present and what effect it may have on the comparison of the mean yields between A and B may be illustrated by using the full formula for the probable error of a difference, as shown in Table 82, page 309. It may be stated that this complete formula is applicable to either a sum or difference.

We may substitute the values from Table 82 in the formula just developed and obtain a value for $\sigma^2_{A\text{-}B}$. The sum of the squares of the deviations of the yields of A from the mean of A, when divided by the number of observations, 10, gives 22.2, and likewise the sum of the squares of the deviations of the yields of B from the mean of B, divided by the number of observations, gives 26.8. The sum of the values in the column $D_A D_B$ is 204, which divided by 10 is 20.4. Summing the mean square values for A and B and subtracting twice the product of $D_A D_B$ we have

$$\sigma^2_{A\text{-}B} = 8.2000$$

and

$$\sigma_{A\text{-}B} = \sqrt{8.2000} = 2.86$$

TABLE 82

METHOD OF DETERMINING THE PROBABLE ERROR OF A MEAN DIFFERENCE BY ELIMINATING THE EFFECT OF CORRELATION

	VARIETY A	VARIETY B	D_A	D^2_A	D_B	D^2_B	D_AD_B	$A-B$	D_{A-B}	D^2_{A-B}
	38	37	0	0	3	9	0	1	–3	9
	40	37	2	4	3	9	6	3	–1	1
	40	40	2	4	6	36	12	0	–4	16
	42	40	4	16	6	36	24	2	–2	4
	39	32	1	1	– 2	4	– 2	7	3	9
	35	30	– 3	9	– 4	16	12	5	1	1
	32	31	– 6	36	– 3	9	18	1	–3	9
	28	22	–10	100	–12	144	120	6	2	4
	42	36	4	16	2	4	8	6	2	4
	44	35	6	36	1	1	6	9	5	25
TOTAL	380	340		222		268	204	40		82
DIVIDED BY N (10)	38	34		22.2		26.8	20.4	4		8.2000

$$\sigma^2_{A-B}=\sigma^2_A+\sigma^2_B-2r\ (\sigma_A\sigma_B)=22.2+26.8-2\ (20.4)=49.0-40.8=8.2$$

$$\sigma_{A-B}=\sqrt{8.2000}=2.86 \quad P.E._{A-B}=\pm.6745\frac{2.86}{\sqrt{N-1}}=\pm.643$$

BESSEL'S FORMULA

$$P.E._M=\pm.6745\sqrt{\frac{\Sigma D^2}{N\,(N-1)}}=\pm.6745\sqrt{\frac{82}{90}}=\pm.6745\sqrt{.911\dot{1}\dot{1}}=\pm.6745\times.955=\pm.644$$

This is the standard deviation of the difference between A and B after the effect of the correlation has been eliminated. The probable error of the difference between the means is given by

$$\pm .6745\frac{2.86}{\sqrt{N-1}}=\pm .643$$

It may be pointed out that the same result for the probable error of the difference between A and B may be obtained from

$$\Sigma D^2_{A\text{-}B}=\Sigma D^2_A+\Sigma D^2_B-\Sigma 2D_AD_B$$

directly. Substituting the necessary values from Table 82 in this formula, we have

$$\Sigma D^2_{A\text{-}B}=222+268-2\ (204)=82$$

Since 82 is the sum of the squares of $D_{A\text{-}B}$, this value may be substituted in the formula for the probable error of the mean

$$\pm .6745\sqrt{\frac{82}{90}}=\pm .644$$

The difference between this probable error and the one obtained above is due to the number of decimals retained.

This is the probable error of the difference when the effect of correlation is removed. It may be noted that it is not necessary to determine the correlation coefficient but merely to eliminate the effect of correlation.

This value for the probable error of the difference between the means, .644, may be compared with the probable error of the difference, 1.57, obtained from the shorter formula, $\sqrt{E_1^2+E_2^2}$, which does not eliminate the effect of correlation. The difference, 4 bushels, divided by the probable error, .64 (read to two decimals), gives a $\frac{D}{P.E.}$ ratio of 6.25. The odds calculated from Table VI for this $\frac{D}{P.E.}$ ratio are over 40,000 to 1, while by using the shorter formula the odds are only 10.70 to 1. This indicates that after eliminating the correlation the difference between A and B is highly significant.

Since the application of the full formula is for the purpose of eliminating the effect of correlation, this may be done by the more convenient method of taking the differences between A and B pair by pair, giving the results in the column *A-B* of Table 82. It should be pointed out that care must be taken to observe the signs. Summing the differences of *A-B* and dividing by the number of observations, 10, the mean difference is obtained, and this is naturally the same as the difference between the two means. The deviations between this mean and the values for *A-B* are obtained and recorded in the column $D_{A\text{-}B}$. These values are squared and summed, giving 82, which is $\Sigma D^2{}_{A\text{-}B}$ and the same value as obtained from the longer formula. This value may be substituted in Bessel's formula and the probable error of the mean obtained. This is a simpler process than the longer formula.

The values in Table 82 may be used to prove that

$$D_{A\text{-}B} = D_A - D_B$$

which relationship was assumed as the first step in the development of the complete formula for the probable error of a sum or difference. Taking the first pair of varieties, we find that $D_{A\text{-}B} = -3$. If we take $D_A - D_B$ we have $0-3$, or -3. For the next pair of varieties we have $D_{A\text{-}B} = -1$. Now $D_A = 2$ and $D_B = 3$, hence $D_A - D_B = 2-3 = -1$. The other values may be similarly compared, proving that $D_{A\text{-}B}$ is equal to $D_A - D_B$.

We may now illustrate how it is possible to combine the measurements of several trials or of several experiments and evaluate the result on the basis of the probable error. Suppose that two varieties of grain A and B have been tested by growing ten plots of each and we are interested in learning whether A is better than B. The yields are

A = 66.6 bushels
B = 61.3 bushels
Gain of A over B = 5.3 bushels

Repeated trials have shown that the probable error of a single plot of the kind used for this experiment is about 12 per cent. We may now use this probable error to compare the results from this experiment.

Since the probable error of the mean of N results equals $\frac{P.E.,}{\sqrt{N}}$, and since in this experiment there were 10 plots of each variety, we have for the value of $P.E._M \frac{12}{\sqrt{10}}$, or 3.80 per cent. That is, the probable error of either one of these means is 3.80 per cent of the mean value. However, we are concerned with the probable error of the difference only, and we obtain the probable error of this difference by multiplying the probable error of the mean in per cent by $\sqrt{2}$, since we are comparing differences. Thus

$$3.80\sqrt{2}=5.37$$

which is the probable error of the difference.

The difference between A and B may now be expressed as the per cent of gain that A shows over B by dividing the difference, 5.3, by the yield of B, 61.3, giving 8.65 per cent. The $\frac{D}{P.E.}$ ratio, or 8.65/5.37, is 1.61. From the table of odds in the Appendix we see that the odds are less than 3 to 1. This indicates that from this experiment we can conclude that there is no significant difference between A and B.

The experiment was repeated and the results from nine plots of each variety were obtained. The yields are

A = 75.0 bushels
B = 69.4 bushels
Gain of A over B = 5.6 bushels

This gain of 5.6 bushels may be expressed in per cent by dividing the gain by the yield of B, or $5.6/69.4=8.07$ per cent. Using the same probable error for a single experiment, 12 per cent, and obtaining the probable error of the difference, we have

$$12/\sqrt{9}\times\sqrt{2}=5.66$$

as the probable error of the difference. With a gain of 8.07 per cent and a probable error of 5.66 the $\frac{D}{P.E.}$ ratio is less than 2 and the odds are too low to denote any significant difference between A and B from this experiment.

It is possible to combine the results of these two experiments in the following way. The first experiment, in which 10 plots of each variety were grown, shows a gain for A over B of 8.65 per cent, and the second experiment, in which 9 plots of each variety were grown, shows a gain for A over B of 8.07 per cent. It is necessary to obtain the weighted average of these percentages and we have

$$\begin{aligned} 8.65\times10 &= 86.50 \\ 8.07\times\ 9 &= 72.63 \end{aligned}$$

Summing

$$19=159.13$$

Dividing 159.13 by the number of plots, 19, we have a gain of 8.38 per cent as the result of the two experiments.

The probable error of this value may be obtained. The probable error of a single plot is 12 per cent, and the probable error of 19 such plots is $12/\sqrt{19}$, or 2.75. Since we are dealing with differences the probable error of the difference is $2.75\sqrt{2}$, or 3.89. The $\frac{D}{P.E.}$ ratio may be obtained by dividing the gain in per cent, 8.38, by the probable error of the difference, 3.89, giving 2.15. Referring to the table of odds in the Appendix we find that the odds are 5.8 to 1. By combining the two experiments we obtain odds that are slightly higher than those obtained with the original experiments, but we do not yet have sufficient information to state whether there is any real difference between A and B.

The experiment was again repeated, and the results from eight plots of each variety are

A = 28.8 bushels
B = 24.8 bushels
Gain of A over B = 4.0 bushels

Expressed in per cent, this gain is 4.0/24.8, or 16.13 per cent. The probable error of the difference is obtained from $12/\sqrt{8}\times\sqrt{2}$, or 6.00. The $\frac{D}{P.E.}$ ratio for this experiment, or 16.13/6.00, is 2.69, and from the table of odds we find that the odds are 13.58 to 1. Again the odds are not high enough to indicate a significant difference.

The results of the three experiments may now be combined by obtaining the weighted average of the percentage gain and the probable error of the experiment. We have

$$\begin{aligned} 8.65\times10 &= 86.50 \\ 8.07\times\ 9 &= 72.63 \\ 16.13\times\ 8 &= 129.04 \end{aligned}$$

Summing

$$27=288.17$$

Dividing 288.17 by 27, the weighted average gain is 10.67. Since there were 27 plots of each variety, the probable error of the gain, which is the probable error of the gain of A over B, is obtained from $12/\sqrt{27} \times \sqrt{2}$, or 3.27. The $\frac{D}{P.E.}$ ratio is 10.67/3.27, or 3.26, and from the table of odds the odds are found to be over 34 to 1.

By combining the results of the three tests we may conclude that the chances are that A is a better variety than B, and with continued trials we may expect that A will yield better than B. This indicates how the results of several experiments may be combined, and while the individual experiments may not lead to significant results, it is possible by combining them that a significant difference may be obtained, especially if A always shows a gain over B. If in one of the three trials B had given a larger yield than A it is evident that the net gain of the three experiments would have been much less than 10.67 per cent.

Another illustration of the application of the probable error in the interpretation of results may be given. Six trials in a rate of seeding test with wheat gave an average yield of 16.8 bushels when 4 pecks of seed were sown, and a yield of 20.8 bushels when 7 pecks of seed were sown. The problem in this case is to determine whether the 7-peck rate is significantly better than the 4-peck rate.

Experience has shown that the probable error for the kind of plots used in this case is approximately 8 per cent, and for 6 trials we would have as the probable error of the difference $8/\sqrt{6} \times \sqrt{2}$, or 4.61. The difference between the 7-peck rate and the 4-peck rate is $20.8 - 16.8$, or 4 bushels. This is expressed in per cent by obtaining 4/16.8, or 23.81 per cent. This value, 23.81, divided by 4.61, the probable error of the difference, gives a $\frac{D}{P.E.}$ ratio of 5.16. Referring to the table of odds it is seen that the odds are high enough to indicate a significant difference between the two rates of seeding.

In this case, however, it is better to use the net difference, since to obtain the larger yield a larger amount of seed, 7 pecks, was used, or a difference of 3 pecks over the lower rate, and it is more accurate to eliminate this amount from the gain. As 3 pecks equals three-fourths of a bushel, or .75, this amount is subtracted from the gain of the heavier seeding over the lighter seeding, 4 bushels, giving a net gain of 3.25 bushels. This net gain is converted into percentage by dividing 3.25 by 16.8, giving a net increase of 19.35. The $\frac{D}{P.E.}$ ratio, or 19.35/4.61, is 4.20, from which odds of about 215 to 1 are obtained. In either case the odds are high enough to indicate significance. When comparing such results it is usually better to base the conclusion on net returns.

Discussion of Odds. In the previous discussion of odds we have assumed that it is equally possible for the deviations to occur in either direction above or below the mean, or in interpreting the results on the basis of the probable error we have assumed that the deviations in either direction are equally likely to occur. In some cases, however, the nature of the material is such as to render deviations in only one direction. Referring again to the two varieties of grain which were compared in three different experiments, it is noted that in all cases A was better than B. This indicates that in general we may expect that A will be better than B, and when comparing results of this sort we may assume that deviations of A as compared with B will be in one direction only, or in the positive direction. When obtaining odds for comparing A with B, it is therefore better to use the probability tables designed for deviations in only one direction.

The probability table, Table VI in the Appendix, is based on the assumption that deviations above or below the mean are equally likely to occur. In cases like the one just cited we assume that the deviations will be in one direction only and therefore one-half of the area of the curve is taken as certainty. This is illustrated in Figure 31.

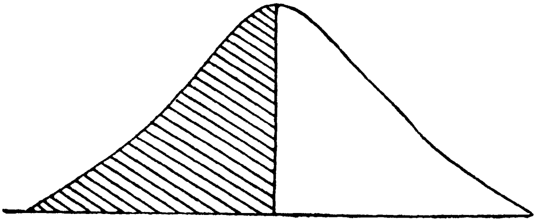


FIG. 31. Illustrating area of curve included for deviations in one direction only.

When using the probability table to obtain odds for results when the deviations are in one direction only we proceed in the following manner. On the basis that the area of the curve used in the table of probability values is 100000, we would add to the values in this probability table one-half of the area of the curve, or 50000. Thus, for $\frac{x}{\sigma} = .60$, we would add 50000 to the probability value in the first column, 22575, giving 72575, and likewise for the other values in the table. Tables of probability values are often presented on the basis of deviations in one direction only, as for example Table II in Pearson's Tables. When 50000 is added to each value in Table VI in the Appendix, the resulting values are comparable with those in Pearson's Table, with the exception that in Table VI the values are read to five numbers only, while in Pearson's Table the area is assumed as 1 and the calculations are carried to several decimals.

For a $\frac{D}{P.E.}$ ratio of 1, this is multiplied by the constant .6745 and we find from the probability table that the area for this value is 75000, as illustrated in Figure 32.

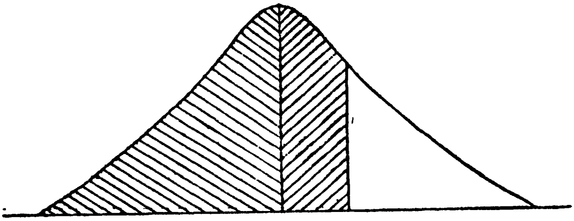


FIG. 32. Illustrating area of curve included for deviations in one direction only when a value equal to $1 \times P.E._s$ is measured from the mean.

To complete the calculation of odds, 75000 is subtracted from the total area of the curve, 100000, giving 25000. Dividing 75000 by 25000 gives odds of 3 to 1. When assuming that the deviations might be in either direction it was found that for a $\frac{D}{P.E.}$ ratio of 1 the odds are 1 to 1, while on the basis that the deviations may be in one direction only the odds are 3 to 1.

Odds for deviations in one direction only for different values of $\frac{D}{P.E.}$ have been calculated and are given in Table 83.

TABLE 83

TABLE OF ODDS TO BE USED WHEN DIFFERENCE IS IN ONE DIRECTION ONLY

DIFFERENCE FROM THE MEAN IN TERMS OF THE PROBABLE ERROR	DIFFERENCE BETWEEN TWO RESULTS IN TERMS OF THE PROBABLE ERROR OF EACH RESULT	ODDS AGAINST SUCH DIFFERENCE OCCURRING DUE TO CHANCE
1.00	1.41	3.0 : 1
1.25	1.77	4.0 : 1
1.50	2.12	5.4 : 1
1.75	2.47	7.4 : 1
2.00	2.83	10.3 : 1
2.20	3.11	13.5 : 1
2.40	3.39	18.0 : 1
2.60	3.68	24.2 : 1
2.80	3.96	32.9 : 1
3.00	4.24	45.5 : 1
3.20	4.52	63.7 : 1
3.40	4.81	90.6 : 1
3.60	5.09	130.8 : 1
3.80	5.37	191.7 : 1
4.00	5.66	285.5 : 1
4.50	6.36	832.3 : 1
5.00	7.07	2630.6 : 1
5.50	7.78	9999.0 : 1
6.00	8.48	33332.3 : 1

When using this table for practical work it is important to be sure that the results being studied are such that odds in one direction only may be expected. For example, when from previous experiments it is known that a certain application of fertilizer gives a gain, comparisons between plots receiving this fertilizer and plots not receiving it are certain to give deviations in one direction only, and therefore odds based on a difference in one direction only should be used. However, in comparing the results from two kinds of fertilizers, as for example two kinds of nitrogen carriers, we may not have previous knowledge that will enable us to judge as to the nature of the difference, and in such cases it is safer to use a table of odds applicable to differences in both directions.

The difference between the odds obtained under the two conditions is that for the same $\frac{D}{P.E.}$ ratio the odds are higher when reading them from a table for deviations in one direction only than they are when read from a table for deviations in both directions. Thus, for a $\frac{D}{P.E.}$ ratio of 2, the odds are 10.3 to 1 on the basis of deviations in one direction only, while on the basis of deviations in both directions the odds are 4.64 to 1.

As a guide for practical application it may be well to state that when in doubt as to which table to use it is better to use the table based on deviations in both directions. This will give lower odds, and will help in safeguarding the conclusions. It is evident that after repeated trials if one treatment, or one variety, tends to give higher values, it is safe to calculate the odds on the basis of deviations in one direction only.

To Determine the Required Number of Observations. In discussing the relation between the probable error of a single determination and the probable error of the mean for any number of observations, it was shown that the probable error of the mean of any number of determinations equals $\frac{P.E._s}{\sqrt{N}}$. Application of this relation may be made to determine the number of individuals, such as the number of plots or animals or any other observation, necessary to obtain the desired degree of accuracy.

For example, we may assume that the probable error for a single small plot is 12 per cent. The probable error of 9 such plots would therefore be $12/\sqrt{9}$, or 4 per cent. Now, how many plots are necessary, assuming that any number can be obtained under similar environmental conditions, to obtain a probable error of the mean of 3 per cent? In this case we are to find the number of plots, which may be derived from the relation $12/\sqrt{N} = 3$. We have

$$12 = 3\sqrt{N}$$

Transposing and squaring

$$9N = 144$$
$$N = 16 \text{ plots}$$

Thus, assuming that the conditions of the experiment will remain the same and that the larger number of plots will not affect the probable error of a single determination, we find that with 16 such plots we would expect the probable error of the mean to be 3 per cent.

This idea may be extended so as to obtain the number of observations necessary to measure differences that may be found between two treatments, two varieties of grain, two different rations in animal feeding, and the like. We may again refer to the experiment comparing the yields of two varieties of grain. The first experiment gave a mean yield for A of 66.6 bushels and a mean yield for B of 61.3 bushels, and from these A showed a gain of 8.65 per cent over B. The result from this experiment was not significant.

We may now determine how many similar plots grown under the same conditions would be necessary to measure this difference and have the results statistically significant. The probable error of a single plot is assumed to be 12 per cent and the difference to be measured is 8.65 per cent. In order that the result will be significant it is necessary to assume odds that may be taken as indicating statistical significance, and odds of 30 to 1 may be accepted. Referring to Table 74 we find the odds nearest to this significance are 31.4 to 1, and working from these odds we find in column 2 a value of 4.52. This value, which is a ratio value of $\frac{D}{P.E.}$, is to be used when two results are being compared on the basis of the probable error of each result. In order to determine the number of observations it is necessary that $\frac{P.E._s}{\sqrt{N}}$ shall equal the ratio of the difference to be measured, in per cent, to 4.52 in order that the odds may be 31.4 to 1. We may arrange the formula as

$$\frac{P.E._s \text{ in per cent}}{\sqrt{N}}=\frac{\text{Percentage difference to be measured}}{\text{Appropriate constant selected to give desired odds}}$$

Substituting the known values, we have

$$\frac{12}{\sqrt{N}}=\frac{8.65}{4.52}$$

Solving

$$12=1.91\sqrt{N}$$

Transposing and squaring

$$3.65N=144$$

from which

$$N=39.45$$

This indicates that we should use about 40 plots of both A and B in order to measure the difference between A and B which has been calculated on the basis of the first experiment. This assumes that the same conditions will hold for the new series of plots as for the first series.

It will be recalled that by repeating the experiment a significant difference was obtained from 27 plots of each variety. That a significant difference was obtained from a smaller number of plots is due to the fact that the same relative comparison between A and B did not hold for all three trials. That is, in the third trial there was a gain of A over B of 16.13 per cent, instead of only 8.65 per cent.

Let us now apply the probable error analysis to a feeding experiment. In this trial four animals were used in each lot. The mean daily gains with their probable errors have been obtained and the results for four of the lots are given here.

LOT	RATION	MEAN DAILY GAIN AND PROBABLE ERROR
I	Clover hay and corn silage	$2.29 \pm .05$
II	Corn silage	$2.32 \pm .06$
III	Sweet clover hay and corn silage	$2.45 \pm .18$
IV	Alfalfa hay and corn silage	$2.49 \pm .05$

One of the purposes of the experiment was to determine the effect of adding leguminous forage to the basal ration, corn silage. The animals in lot II were given the basal ration of corn silage and the animals in lots I, III, and IV received leguminous hay in addition to corn silage. Comparing lots IV and II we have

Lot IV	$2.49 \pm .05$
Lot II	$2.32 \pm .06$
Gain of IV over II	$.17 \pm .08$

The $\frac{D}{P.E.}$ ratio is .17/.08, or 2.12, and is not high enough to denote significance. It is evident that the other two lots, I and III, when compared with II would show no significant gain for the leguminous hay.

We may now determine how many animals would be needed in such an experiment to be able to measure the differences obtained and have a significant result. In addition to the data on the four lots given here, data are available on six other lots, and from these ten lots the probable error of a single animal is obtained in the following manner. The daily gain for each experiment is summed, giving 22.35. For each lot the probable error of a single animal is obtained by multiplying the probable error of the mean by $\sqrt{4}$, since there were four animals in each lot. Summing the probable errors obtained in this way for each of the ten lots we have a total of 1.88. Dividing this by the total daily gain, 22.35, gives the probable error in per cent for a single animal, or $P.E._s = 8.41$. The gain of lot IV over lot II may be expressed in per cent by dividing the gain, .17, by the mean of lot II, 2.32. This gives 7.33 as the percentage gain.

Substituting the necessary values in the general formula for determining the number of individuals and assuming that we require odds of at least 30 to 1, we have

$$\frac{8.41}{\sqrt{N}} = \frac{7.33}{4.52}$$

Solving

$$8.41 = 1.62\sqrt{N}$$

Transposing and squaring

$$2.62N = 70.73$$

from which

$$N = 27.00$$

This shows that for a difference of only 7.33 per cent 27 animals would have to be used in each lot. It would be difficult to measure such a small difference as this under ordinary conditions. If several experiments with a smaller number of animals all showed similar results, perhaps by combining them it would be possible to obtain significant results.

It is well to point out that in making application of this formula to determine the number of observations necessary to measure a difference it must be understood that it is on the basis that the percentage difference to be measured will remain the same and that the probable error of a single determination will not differ from the value used in the equation. If the conditions are such that the probable error of a single plot should be less when the calculated number of plots are grown than it was assumed to be, and if the percentage difference should remain the same, then it would not be necessary to have the number of plots as calculated. Again, it is possible that when a large number of observations are needed it may be necessary to extend the experimental area to such an extent that the probable error of a single plot may be increased. In extending the area to give the number of plots it may be necessary to include a soil of a decidedly different fertility value, and the probable error of a single plot may be increased unless some method of arrangement is followed that will tend to eliminate this difference in fertility.

When making use of this formula to determine the number of plots or observations necessary it may be suggested that to be on the safe side it is well to add a few additional plots. For example, where it was found that approximately 40 plots are needed it may be well, both for safety and for accuracy, to have perhaps 45 plots. The question may arise as to whether it is necessary to grow all of these plots the same year or under the same conditions. In answer to this it may be stated that there is nothing in the formula that requires that all of the plots be grown in one year, or in other words we may say that mathematically it is not necessary. From a practical standpoint it is better to do so, since environmental factors will likely not be the same should the 40 or 45 plots be divided into three different groups or grown in three different years. Therefore, while mathematically it makes no difference whether or not the plots are all grown at one time, from the practical standpoint if it is necessary to divide them or grow them in different years, then a larger number of plots should be grown than required from the formula. In other words, when 40 plots are necessary and if it is impossible to have the entire number in any one year, it is better from the standpoint of practical field experiments to grow about 16 plots each year for three years. That is, the application of the formula making it possible to determine the number of plots necessary assumes that they may be grown under the same conditions, or as nearly as possible the same conditions, as the original experiment, and since fluctuations will occur it is more desirable to have a larger number to offset such fluctuations.

In using this formula we have taken 4.52 as the ratio value to lead to odds of 31.4 to 1. This is on the assumption that we are comparing the results on the basis of the probable error of each result. If desired the probable error of the difference may be substituted in the formula in place of the probable error of a single plot. The probable error for the plots, used in comparing varieties A and B above, is 12 per cent, and for the probable error of the difference we have

$$12\sqrt{2}=16.97 \text{ per cent}$$

This will be substituted in the numerator of the formula in place of the probable error of a single plot. Then the factor which will lead to odds of 31.4 to 1 is taken from the first column of Table 74, since we now are dealing with the probable error of the difference. This factor is 3.2. Substituting these values in the formula we have

$$\frac{16.97}{\sqrt{N}}=\frac{8.65}{3.2}$$

Solving

$$16.97=2.70\sqrt{N}$$

Transposing and squaring

$$7.29N=287.98$$

from which

$$N= 39.50$$

This formula leads to the same number of plots required as determined from the probable error of a single plot.

In using this formula, if the probable error of a single plot is being considered, the ratio value should be taken from the second column of Table 74. If the probable error of the difference between two plots is used, the ratio value should be obtained from the first column of Table 74.

If an experiment is planned for the purpose of comparing certain varieties or treatments in which it is definitely known that one variety or treatment will give a larger return than another, the values in Table 83 may be used, but it should be stated that it is necessary to know definitely that one result will be better than another. Again, it may be pointed out that to be on the safe side the values in Table 74 should be used. The result when using the latter values is that more observations are required and therefore there is likely to be a sufficient number for odds in both directions or in either direction.

Other Methods of Comparing Results. It will be recalled that in the comparison of the two varieties of wheat in Table 82 before and after eliminating the effect of correlation, odds were obtained which were very different for the two comparisons. The comparison without eliminating the effect of the correlation showed that there was no difference between A and B. After the correlation was eliminated odds of over 40000 to 1 were obtained, which would be interpreted to mean that on repeated trials we would expect A to be better than B, and that the difference is not due to chance variation but to a real difference in the yielding ability of the two varieties.

The question may arise as to whether such large odds are dependable. It will be recalled that these odds were obtained from the normal probability table, or the table of odds based on the values in the probability table, and that these in turn are based on the assumption of a normal distribution for the observations on which the result is determined. In other words, we have assumed that the standard deviation is known exactly, but investigations have shown that for small samples the true standard deviation can only be roughly approximated from the sample at hand 'Student' was one of the first to emphasize this fact, and he pointed out that for small samples while a value s may be a perfectly good estimate of σ it seldom or ever equals σ. While σ may be distributed in accordance with the normal curve we cannot assume that s will be so distributed, or that the results from observations based on small samples can be tested by the normal probability tables. 'Student' investigated this problem and studied the distribution of s or s^2 in random samples. As a result of his studies he developed a set of probability values to be used in interpreting results obtained from small samples.

When interpreting results on the basis of these values the mean and standard deviation of the sample are determined and the ratio of the mean to the standard deviation, which ratio 'Student' has called Z, is obtained. From this ratio and the number in the sample reference is made to the special probability table developed by 'Student' and the probability determined. From this probability the odds may be calculated in the usual way, and these odds indicate whether the mean may have arisen due to chance variation or whether it may have been affected by some special factor or factors. If the odds are low, we conclude that the result may have been due to chance. If the odds are high, we conclude that there is a real difference between the varieties or treatments. In connection with experiments in biology and agronomy 'Student's' table has been found to be valuable, and a convenient application has been to determine first the differences which may exist between a series of A and a series of B values pair by pair, but the pairing is not to be considered as a part of 'Student's' method.

We may apply this method of interpreting the difference between variety A and variety B from Table 82. The necessary steps are followed in Table 84, page 326. The differences between A and B are obtained pair by pair and the mean of this series of differences is determined, as was done in Table 82. By obtaining the deviations of the differences from this mean the standard deviation or standard error of this mean is determined. The sum of the squared deviations is 82, and dividing this by the number of observations, 10, and extracting the square root we have 2.86 as the standard deviation.

In using 'Student's' probability values to interpret this result the ratio of the mean to its standard deviation is obtained, and is designated Z. This may be interpreted by referring to Table VIII in the Appendix. This table has been prepared from 'Student's' table by calculating the odds directly for different values of Z and N. For convenience additional values are included in this table of odds. These additional values were obtained by direct interpolation and are satisfactory for practical use. The use of the table may now be illustrated.

For the data in Table 84 we find the Z value to be 1.40, and with $N = 10$ we find from Table VIII that the odds are 908 to 1.

TABLE 84

APPLICATION OF 'STUDENT'S' METHOD OF INTERPRETATION TO THE DIFFERENCE BETWEEN TWO VARIETIES OF GRAIN

VARIETY A	VARIETY B	$A-B$ or D	D_{A-B}	D^2_{A-B}
38	37	1	–3	9
40	37	3	–1	1
40	40	0	–4	16
42	40	2	–2	4
39	32	7	3	9
35	30	5	1	1
32	31	1	–3	9
28	22	6	2	4
42	36	6	2	4
44	35	9	5	25
		10)40		10) 82
	MEAN	=4.00		8.2000
				$\sigma=\sqrt{8.2000}=2.86$

$$Z=\frac{M}{\sigma}=\frac{4.00}{2.86}=1.40$$

For N =10 Odds 908 to 1

It is to be noted that while these odds are still high enough to indicate a significant difference between A and B, they are much lower than those obtained from the normal probability values. For the comparison of small samples, therefore, it is better to use 'Student's' probability values rather than the normal probability values. As the size of the sample increases there is a closer relationship between the values obtained from 'Student's' table and from the normal probability table, and for N beyond 30 the normal probability values may be used in place of 'Student's' values with slight modification.

As an illustration of the use of 'Student's' method for cases where N is greater than 30 we will use an example in which $N=30$, in order to make a comparison between the results obtained from 'Student's' probability values and from the normal probability values. These data are from a comparison of two varieties of wheat, each grown on 30 plots. The differences were obtained and the mean difference and its standard deviation were found to be 30.9667 and 57.7878. 'Student' has not tabled probabilities for N greater than 30, but to illustrate the method to be followed for such cases we proceed as follows. The standard deviation is divided by $\sqrt{N-3}$ and the ratio of the mean to this value is obtained. Thus

$$\frac{57.7878}{\sqrt{30-3}}=11.1212 \text{ and } \frac{30.9667}{11.1212}=2.78$$

With this ratio value the normal probability tables (Table VI in the Appendix) are used and the odds are 367 to 1. For comparison we may use 'Student's' table and find the odds to be 242 to 1. The odds from the normal probability table are a little higher than those obtained from 'Student's' table but the difference is not so great but that this method may be used when N is greater than 30. In either case the odds are high enough to indicate significance.

TABLE 85

APPLICATION OF 'STUDENT'S' METHOD OF INTERPRETATION TO A RATE OF SEEDING EXPERIMENT

YIELD FROM 9 PECKS PER ACRE A	YIELD FROM 6 PECKS PER ACRE B	$A-B$	D_{A-B}	D^2_{A-B}
21.6	19.9	1.7	.6	.36
25.7	23.6	2.1	1.0	1.00
38.6	35.2	3.4	2.3	5.29
16.8	14.6	2.2	1.1	1.21
31.8	32.2	− .4	−1.5	2.25
41.4	41.0	.4	− .7	.49
33.5	35.4	−1.9	−3.0	9.00
		7)7.5		7)19.60
		MEAN = 1.1		2.80

$$\sigma=\sqrt{2.80}=1.7$$

$$Z=\frac{1.1}{1.7}=.65$$

For N =7 Odds approximately 11 to 1

The use of 'Student's' method of interpretation may be applied to an experiment in rate of seeding with wheat. The tests were conducted for several years and the results for two rates of seeding are given in Table 85, page 327. Obtaining the gain of the 9–peck rate over the 6–peck rate and calculating the standard deviation and Z, from Table VIII the odds are found to be approximately 11 to 1 for $N=7$. In this case the odds are not high enough to indicate that the increased rate of seeding has resulted in a decided gain. This interpretation is on the basis of the actual yields, but from a strictly agronomic standpoint it is important to know how much net increase is obtained by using more seed. Since 9 pecks are 3 pecks, or .75 of a bushel, more than 6 pecks, this amount may be subtracted from the difference, giving the net difference for the increased rate of seeding. This net difference divided by the standard deviation will give lower odds. It may be stated that the net difference when subtracted from the mean difference gives the same result as would be obtained if the net difference were subtracted from each individual experiment. In other words, since we are subtracting a constant it makes no difference whether it is subtracted from each individual difference between A and B or whether it is subtracted from the average difference.

After developing the probability values for use with argument Z, 'Student' considered the question further and developed another set of values for argument t, in which $t=Z\sqrt{N-1}$. For convenience the probability values in this table have been recalculated in the Department of Plant Breeding at Cornell University, so that the odds may be read directly, and appear as Table IX in the Appendix. The degrees of freedom, taken as one less than the number of items, are given at the top of the table and the t values appear in the first column.

The interpretation of results by means of the values in this table may now be illustrated with the data in Table 84. In Table 84 Z was found to be 1.40, and since there are 10 observations t may be obtained from $Z\sqrt{N-1}$, or $t=1.40\sqrt{9}$, or 4.20. Referring to Table IX for t equalling 4.20, and 9 degrees of freedom, we have odds of 832 to 1. This gives odds very similar to those obtained when using the value for Z and $N=10$.

If it is preferred to determine t directly in an example of this sort, we first obtain the standard error of the mean difference from

$$\sigma_{M_{A-B}} = \sqrt{\frac{\Sigma D^2_{A-B}}{N\,(N-1)}}$$

Substituting the value for ΣD^2_{A-B}, 82, and the value for N from Table 84, we find the standard error to be .95. Then t is obtained from

$$t = \frac{M}{\sigma_{M_{A-B}}} = \frac{4.00}{.95} = 4.21$$

This is approximately the same value as obtained from $Z\sqrt{N-1}$.

In this problem we have considered that there is a relation between the plots of A and the plots of B, and that we were justified in obtaining the differences pair by pair. There are many cases in

TABLE 86

APPLICATION OF 'STUDENT'S' t METHOD FOR DETERMINING THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO MEANS

VARIETY A	VARIETY B	D_A	D_B	D^2_A	D^2_B
38	37	0	3	0	9
40	37	2	3	4	9
40	40	2	6	4	36
42	40	4	6	16	36
39	32	1	− 2	1	4
35	30	− 3	− 4	9	16
32	31	− 6	− 3	36	9
28	22	−10	−12	100	144
42	36	4	2	16	4
44	35	6	1	36	1
380	340			222	268

$N_1 = 10$ $N_2 = 10$

$$M_A = \frac{\Sigma A}{N_1} = \frac{380}{10} = 38 \qquad M_B = \frac{\Sigma B}{N_2} = \frac{340}{10} = 34$$

$$t = \frac{(M_A - M_B)\sqrt{N_1 + N_2 - 2}}{\sqrt{\Sigma D^2_A + \Sigma D^2_B}} \sqrt{\frac{N_1 N_2}{N_1 + N_2}}$$

$$t = \frac{38 - 34\sqrt{10+10-2}}{\sqrt{222+268}} \sqrt{\frac{(10)\,(10)}{10+10}} = \frac{4\sqrt{18}}{\sqrt{490}} \sqrt{\frac{100}{20}} = 1.72$$

Degrees of freedom $= N_1 + N_2 - 2 = 18$

Odds approximately 18 to 1

which there is no relation between the measurements of one series and the measurements of another and therefore we are not justified in obtaining the paired differences, but we are interested in determining whether there is any significant difference between the means. In such cases we proceed in a different manner. We may use the same data as in Table 84 and determine first the means of A and B. The deviations of the individual results from their respective means are obtained, squared, and summed, giving the values for ΣD^2_A and ΣD^2_B, as indicated in Table 86, page 329.

We may obtain t from the formula

$$t=\frac{(M_A-M_B)\sqrt{N_1+N_2-2}}{\sqrt{\Sigma D^2_A+\Sigma D^2_B}}\sqrt{\frac{N_1N_2}{N_1+N_2}}$$

In this formula N_1 and N_2 are the number of observations in each case. Substituting the values from Table 86 we find t to be 1.72. It may be pointed out, however, that when there are an equal number of items in both series it may be more convenient to obtain t by first determining the standard error of the difference, σ_D, between the two means from

$$\sigma_D=\sqrt{\frac{\Sigma D^2_A}{N\,(N-1)}+\frac{\Sigma D^2_B}{N\,(N-1)}}$$

Substituting the values from Table 86 in this formula we have

$$\sigma_D=\sqrt{\frac{222}{90}+\frac{268}{90}}=2.33$$

and

$$t=\frac{M_D}{\sigma_D}=\frac{4.00}{2.33}=1.72$$

The degrees of freedom are $N_1+N_2-2=18$, in which N_1 and N_2 refer to the number of observations in each case. From Table IX we find for $t=1.72$ and for 18 degrees of freedom that the odds are about 18 to 1.

We would conclude from this method of analysis that there is no significant difference between A and B. However, from 'Student's' Z test and from the first value of t obtained, in which it is assumed that there is a relation between the results for A and B, we obtain odds high enough to indicate significance. This indicates that sometimes one of the methods may indicate significant differences and the others may not, and in such cases this fact cannot be ignored even though all methods do not show significance. It must be understood that the first two methods need not always be applied to paired results, because they may be used with a single series of observations and will then show whether the mean differs significantly from zero.

Fisher has also considered the problem of small samples, and has prepared a table for interpreting results on the basis of t. It is not possible to present Fisher's complete table, but in the last column of Table X, which is taken from Snedecor, part of the values from Fisher's table are given. The values of t given in this table are those indicating odds of 19 to 1 and 99 to 1, since we are chiefly concerned in knowing whether or not our results are significant. The numbers in light-face type are those indicating odds of 19 to 1 and those in dark-face type indicate odds of 99 to 1. Any value for t lower than the values tabled for the degrees of freedom concerned would not be considered significant. The degrees of freedom are given in the first column at the left of the table. It should be pointed out that 'Student's' t values are calculated on the basis of deviations in one direction only, while Fisher's t values are for deviations in both directions.

Fisher has suggested the following method for determining t:

$$t=\frac{M\sqrt{N}}{s}$$

where M is the deviation of the sample mean from the population mean, N is the number of observations, and s is the square root of the variance, obtained by dividing the sum of the squares of the deviations by the degrees of freedom. Degrees of freedom are taken as 1 less than the number of observed values.

We may use the data from Table 84 to show the application of Fisher's method, as given in Table 87.

TABLE 87
APPLICATION OF FISHER'S METHOD FOR DETERMINING THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO MEANS

VARIETY A	VARIETY B	$A-B$ or x	$x-M$	$(x-M)^2$
38	37	1	−3	9
40	37	3	−1	1
40	40	0	−4	16
42	40	2	−2	4
39	32	7	3	9
35	30	5	1	1
32	31	1	−3	9
28	22	6	2	4
42	36	6	2	4
44	35	9	5	25
		$\Sigma x=40$		$\Sigma(x-M)^2=82$

$$N=10 \qquad M=\frac{\Sigma(x)}{N}=\frac{40}{10}=4$$

$$\frac{s^2}{N}=\frac{\Sigma\,(x-M)^2}{N\,(N-1)}=\frac{82}{(10)\,(9)}=.911111$$

$$s=\sqrt{.911111}\sqrt{10}=(.955)\ (3.162)=3.020$$

or

$$s=\sqrt{\frac{\Sigma\,(x-M)^2}{N-1}}=\sqrt{\frac{82}{9}}=3.018$$

$$t=\frac{M\sqrt{N}}{s}=\frac{4\sqrt{10}}{3.018}=\frac{12.648}{3.018}=4.19$$

or

$$t=\frac{M}{\sqrt{\frac{s^2}{N}}}=\frac{4}{\sqrt{.911111}}=4.19$$

Degrees of freedom $=N-1=9$
Odds more than 99 to 1

The values needed are M, s, and t, and they are obtained from the following formulas.

$$M=\frac{\Sigma\,(x)}{N}$$

$$\frac{s^2}{N}=\frac{\Sigma\,(x-M)^2}{N\,(N-1)}\text{ or }s^2=\frac{\Sigma\,(x-M)^2N}{N\,(N-1)}\text{ or }\frac{\Sigma\,(x-M)^2}{N-1}$$

$$t=\frac{M\sqrt{N}}{s}$$

In these formulas M and s have already been explained, N refers to the number of observations, and x to the deviations between the individual observations. The degrees of freedom are $N-1$.

Substituting the values from Table 87 we find $t=4.19$. This is practically the same value for t as obtained from $Z\sqrt{N-1}$. From Table X, for 9 degrees of freedom, we find that a t value of 3.250 due to chance alone may be expected once out of 100 trials. From the value obtained for t, 4.19, we may be certain that the odds are more than 99 to 1 that there is a significant difference between varieties A and B. This method may also be used to determine whether an observed mean differs significantly from zero.

In applying this method we have assumed that correlation exists and that we are justified in obtaining the paired differences between A and B, but often it is important to determine the significance between means when there does not seem to be any justification for comparing the results pair by pair, or perhaps when due to a

TABLE 88

APPLICATION OF FISHER'S METHOD FOR DETERMINING THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO MEANS WHEN THE RESULTS CANNOT BE PAIRED

VARIETY x_1	VARIETY x_2	x_1-M_1	x_2-M_2	$(x_1-M_1)^2$	$(x_2-M_2)^2$
81	35	27	-16	729	256
48	52	- 6	1	36	1
46	75	- 8	24	64	576
53	43	- 1	- 8	1	64
47	49	- 7	- 2	49	4
52	51	- 2	0	4	0
52		- 2		4	
379	305			887	901

$N'_1=7$ $\qquad$ $N'_2=6$

$$M_1=\frac{\Sigma\ (x_1)}{N'_1}=\frac{379}{7}=54 \qquad M_2=\frac{\Sigma\ (x_2)}{N'_2}=\frac{305}{6}=51$$

$$s^2=\frac{\Sigma\ (x_1-M_1)^2+\Sigma\ (x_2-M_2)^2}{(N'_1-1)+(N'_2-1)}=\frac{887+901}{6+5}=\frac{1788}{11}=162.545455$$

$$s=\sqrt{162.545455}=12.749$$

$$t=\frac{M_1-M_2}{s}\sqrt{\frac{(N'_1)\ (N'_2)}{N'_1+N'_2}}$$

$$t=\frac{54-51}{12.749}\sqrt{\frac{(7)\ (6)}{(7+6)}}=\frac{3}{12.749}\sqrt{\frac{42}{13}}=(.235)\ (1.797)=.422$$

Degrees of freedom are $N=(N'_1-1)+(N'_2-1)=11$
Odds not significant.

different number of observations in each series it is not possible to pair the results. The method has been extended by Fisher for this sort of problem. The formulas are

$$N' = \text{Number observed} \quad M_1 = \frac{\Sigma\,(x_1)}{N'_1} \text{ and } M_2 = \frac{\Sigma\,(x_2)}{N'_2}$$

$$s^2 = \frac{\Sigma(x_1 - M_1)^2 + \Sigma(x_2 - M_2)^2}{(N'_1 - 1) + (N'_2 - 1)}$$

$$t = \frac{M_1 - M_2}{s}\sqrt{\frac{(N'_1)\,(N'_2)}{N'_1 + N'_2}}$$

The degrees of freedom are $N = (N'_1 - 1) + (N'_2 - 1)$

The various steps are followed in Table 88, page 333, and the values substituted in the formulas. The value for t is .422, and referring to Table X for 11 degrees of freedom the lowest value tabled is 2.201. Since the value obtained, .422, is much lower than this, we conclude that there is no significant difference between the two means.

Another example is given in Table 89, showing the comparison of two varieties of grain.

TABLE 89

APPLICATION OF FISHER'S SECOND METHOD IN A COMPARISON OF TWO VARIETIES OF GRAIN

VARIETY x_1	VARIETY x_2	$x_1 - M_1$	$x_2 - M_2$	$(x_1 - M_1)^2$	$(x_2 - M_2)^2$
29.3	20.4	-1.7	-1.4	2.89	1.96
21.3	20.2	-9.7	-1.6	94.09	2.56
30.7	20.1	- .3	-1.7	.09	2.89
30.2	23.4	- .8	1.6	.64	2.56
36.4	23.7	5.4	1.9	29.16	3.61
37.4	19.3	6.4	-2.5	40.96	6.25
35.6	25.1	4.6	3.3	21.16	10.89
27.8	18.7	-3.2	-3.1	10.24	9.61
24.7	17.4	-6.3	-4.4	39.69	19.36
36.3	29.6	5.3	7.8	28.09	60.84
309.7	217.9			267.01	120.53

$N'_1 = 10$ $\qquad$ $N'_2 = 10$

$$M_1 = \frac{309.7}{10} = 31.0 \qquad M_2 = \frac{217.9}{10} = 21.8$$

$$s^2 = \frac{267.01 + 120.53}{9 + 9} = 21.530000 \quad s = \sqrt{21.530000} = 4.640$$

$$t = \frac{31.0 - 21.8}{4.640}\sqrt{\frac{(10)\,(10)}{10 + 10}} = 4.434$$

Degrees of freedom = 18

Odds more than 100 to 1

Following the various steps we have a t value of 4.434, which indicates odds of over 100 to 1. These odds are high enough to indicate a significant difference between the two varieties.

It is evident that since the results may be paired it is possible to use Fisher's first method as described above for the analysis of these data. At times it is desirable to use both methods, especially if the result is considered to be of great importance, for one method may show significance while the other may fail to do so. In such cases some significance would be attached to the result. The point to keep in mind regarding these two methods is that when there is no just reason for expecting a correspondence between results or when one is not justified in pairing the results the second method only should be used. For example, we may have a series of yields for two varieties of grain in which in one case 20 experiments have been reported and in the other case only 15 experiments have been reported. The experiments may not have been conducted on the same fields, and still we may evaluate the results by using the second method of Fisher.

In this chapter different methods have been presented which may be used for the analysis of results from small samples. Experience has shown that it is better to use the methods more recently developed, that is, 'Student's' and Fisher's, rather than the older methods. We may conclude that these newer methods should be used wherever they may be applied.

Further Applications of Probability. Another very useful method for determining the reliability of results or predictions is that developed by Pearson, enabling us to predict future expectations on the basis of past experience. This method is particularly useful for results that are obtained or reported in groups such that the usual methods of calculating the probable error or determining the probability cannot be used.

Suppose that a first sample, N, has been obtained, and that p represents the number of times that a result occurs and q represents the number of times it fails. Then

$$\bar{p}=\frac{p}{N} \text{ and } \bar{q}=\frac{q}{N}$$

$$\bar{p}+\bar{q}=1$$

Now suppose another sample, m, from the same material has been obtained. The expected mean of the second sample is

$$\text{Mean}=m\bar{p}+\frac{m}{N+2}(\bar{q}-\bar{p})$$

The mode equals the integral portion of $m\bar{p}+\bar{p}$ and the standard deviation is obtained from

$$S.D.=\sqrt{m\left(\bar{p}+\frac{\bar{q}-\bar{p}}{N+2}\right)\left(\bar{q}-\frac{\bar{q}-\bar{p}}{N+2}\right)\left(1+\frac{m-1}{N+3}\right)}$$

These formulas may be applied to the analysis of the following data from the third generation of a cross between two varieties of wheat, Marquis and Turkey. On August 9 the data obtained were

	BUNT-FREE	BUNTED OR PARTLY BUNTED	TOTAL
Number of Plants	913	350	1263

On August 16 a total of 3763 plants were harvested. What would be the expected number of bunt-free plants, and is there any difference in the two lots and is this difference greater than would be expected from chance variation?

Let p represent the number of bunt-free plants and q the number of bunted or partly bunted plants. We may determine $\bar{p}$ and $\bar{q}$ from the first results.

$$\bar{p}=\frac{p}{N}=\frac{913}{1263}=.723$$

$$\bar{q}=\frac{q}{N}=\frac{350}{1263}=.277$$

The expected mean in the second series is obtained from the formula above.

$$M = 3763\ (.723) + \frac{3763}{1265}\ (.277 - .723) = 2719.322$$

From the formula above the mode, if needed, is obtained as follows:

$$\text{Mode} = 3763\ (.723) + .723 = 2721.372$$

or

$$\text{Mode} = 2721$$

For the standard deviation

$$S.D. = \sqrt{3763\left(.723 + \frac{-.446}{1265}\right)\left(.277 - \frac{-.446}{1265}\right)\left(1 + \frac{3762}{1266}\right)} = 54.73$$

$$P.E. = \pm .6745\sigma = \pm (.6745)\ (54.73) = \pm 36.92$$

The observed results from the plants harvested on August 16 are

	BUNT-FREE	BUNTED OR PARTLY BUNTED	TOTAL
Number of Plants	3092	671	3763

It is seen that the actual number of bunt-free plants obtained in the second series is 3092, while the expected number as calculated is 2719.322 ± 36.92. It is evident that the difference is real and that there are fewer plants showing bunt among those ripening later.

This method is particularly useful with such problems as arise from experiments in plant pathology and the study of reactions of plants to diseases in different environments. There are many other applications that may also be made.

For determining the probable error of Mendelian results we have the following formulas, in which N is the total number of observations and p and q are explained for each formula.

For ratios

$$P.E. = \pm .6745\sqrt{\frac{p\,q}{N}}$$

For absolute numbers

$$(1)\quad P.E. = \pm .6745\sqrt{p\,q\,N}$$ *(where p and q represent expected percentages of frequency)*

$$(2)\quad P.E. = \pm .6745\sqrt{\frac{\text{Product of expected numbers}}{N}}$$

For percentages

$$P.E. = \pm .6745\sqrt{\frac{(p\,q)\text{ in }\%}{N}}$$

We may apply these formulas in interpreting the results from a test for inheritance of kernel color in wheat. The results from 200 plants show 148 plants with red kernels and 52 plants with white kernels. This indicates a 3 to 1 ratio, and the probable error for this ratio is obtained from the formula

$$P.E. = \pm .6745\sqrt{\frac{p\,q}{N}}$$

Substituting in this formula we have

$$P.E. = \pm .6745\sqrt{\frac{(3)\,(1)}{200}} = \pm .08$$

For the probable error for absolute numbers we may express the 3 to 1 ratio in percentage, and p and q will be .75 and .25. Substituting these values in the formula

$$P.E. = \pm .6745\sqrt{p\,q\,N}$$

we have

$$P.E. = \pm .6745\sqrt{(.75)\,(.25)\,(200)} = \pm 4.13$$

This probable error for absolute numbers may also be obtained from the formula

$$P.E. = \pm .6745\sqrt{\frac{\text{Product of expected numbers}}{N}}$$

For a 3 to 1 ratio for 200 individuals we would expect 150 plants with red kernels to 50 plants with white kernels. Substituting these numbers in the formula we have

$$P.E. = \pm .6745\sqrt{\frac{(150)\,(50)}{200}} = \pm 4.13$$

For the probable error for percentage for a 3 to 1 ratio, we substitute in the formula

$$P.E. = \pm .6745 \sqrt{\frac{(p\,q) \text{ in } \%}{N}}$$

the values of p and q in per cent, or .75 and .25, and have

$$P.E. = \pm .6745 \sqrt{\frac{(.75)\,(.25)}{200}} = \pm .02$$

These probable errors may now be applied to the observed data. For ratios, we find that the observed numbers of 148 plants with red kernels and 52 plants with white kernels give a ratio of 2.96 to 1.04. The probable error is .08. The deviation from the expected 3 to 1 ratio is .04, and since this deviation is only one-half as large as the expected probable error for this ratio the result is significant. The probable errors for numbers and percentages may be compared in the same way and it will be found that similar results are obtained.

The Goodness of Fit, as discussed in Chapter XI, is frequently applied in the interpretation of Mendelian results and is a better measure when it can be used, as it is based on the entire distribution.

CHAPTER XIII

ANALYSIS OF VARIANCE

The methods for the analysis of the results of experiments on the basis of the standard error and the probable error are useful and have their proper place in statistical interpretation. Since these methods are of such nature that they include all of the causes of variation, it is important at times to have some process of analysis that will make possible the elimination of the effect of certain causes, thus affording a more exact analysis of the data and a more correct determination of the real error of the experiment. Fisher has given such a method in the analysis of variance, and he and his coworkers have done much to extend its application and to demonstrate its value in the analysis of the results of many kinds of experiments.

This method of the analysis of variance is based on the fact that the total variation is the resultant of several factors. Some of these factors may be known and it may be possible to eliminate their effect, and this is one of the important contributions of the method. For example, suppose we have a series of plot yields obtained from several plots each sown to a different variety of wheat. The total variation is then made up of several factors, such as the variation between the varieties, the variation due to the blocks or location in the field, and the like, which may be determined and separated from the total variation. The residue, which cannot be attributable to known causes, is the variation due to the error of the experiment. We may therefore think of variance as being an additive quantity.

In order to illustrate this idea, suppose that we represent the total variance by σ^2_T, which is the result of the variances of several components. For a field experiment similar to the one mentioned we have

$$\sigma^2_T = \sigma^2_V + \sigma^2_B + \sigma^2_R \text{ and so on.}$$

In this expression σ^2_V and σ^2_B are the variances for the effect of varieties and blocks and σ^2_R is the residue, or the variance due to error. This illustrates the additive nature of variance, and indicates how the total variance may be separated into its several components. There will always be a certain amount unaccounted for and this is assigned to error.

Fisher has introduced another term, the degrees of freedom, which has accompanied his analysis of variance, and the idea of which may be applied also in other types of analysis. We may think of the degrees of freedom in this way. Suppose that we have 10 measurements, such as 10 plot yields of the same treatment. The mean of the 10 yields may be obtained, and it is assumed that 9 of the observations are free to vary. Thus we have 9 degrees of freedom. In general, degrees of freedom equal the number of observations, or comparisons, less the number of constants that have been determined from these observations. In the case cited, we have determined the mean of 10 observations. Therefore, 1 degree of freedom is lost and there are 9 degrees of freedom remaining.

The different factors affecting the total variation contribute a certain amount not only to the total variation but they also contribute a certain number of degrees of freedom. For example, if we have 5 varieties that are replicated in 5 plots we have 25 plots in all. For the total experiment we have $N-1$, or 24, degrees of freedom for the total variation. Since there are 5 varieties we have $N-1$, or 4, degrees of freedom for varieties, and since there are 5 replicates we have $N-1$, or 4, degrees of freedom for replicates. When the variation due to varieties is eliminated from the total variation we also deduct the 4 degrees of freedom for these varieties, leaving 20 degrees of freedom. If the variation due to the 5 blocks is also eliminated the number of degrees of freedom is likewise reduced by 4, leaving 16 degrees of freedom. If these are all of the known components of the total variation that may be eliminated, then for the effect of the unknown factors we have the remainder after subtracting from the total the amount due to varieties and the amount due to blocks, and the degrees of freedom are likewise obtained by subtraction.

This remainder is thought of as the amount due to error, and the remaining degrees of freedom as the degrees of freedom for error. In general, as the different components are deducted from the total variation there is a reduction in the error, provided that the amounts of the different components taken out are of appreciable size compared with the number of degrees of freedom that are eliminated at the same time.

One of the requirements for the analysis of variance is that the sample analyzed shall really be a random sample of the condition being studied. This implies that if the results are those obtained from field or other kinds of experiments the plots or individuals from which the measurements are taken must have been distributed at random. That is, there should be no systematic arrangement nor any effort to locate a particular plot or individual by other than random or chance location. For example, if we were conducting an experiment with five plots of five treatments or varieties the location of the five plots in the field with reference to each other must be entirely by random. We would not take five plots and arrange the five treatments definitely in systematic order, but the treatment to be assigned to each particular plot will be determined by random, as by the throwing of dice, drawing of numbered cards, or some other chance assignment. We may take five cards and number them separately, and after thorough shuffling draw out one card. This number will indicate the treatment to be placed on plot 1, for instance. If we choose we may have a series of cards with the five treatment numbers and another series of cards with the five plot numbers, and draw one card for plots and one for treatment. For example, if plot number 3 and treatment number 5 are drawn at the same time, it means that treatment 5 is to be placed on plot 3, and so on for the other plots.

It must be understood that the application of the idea of variance analysis is by no means limited to field experiments or experiments of that nature. It is of use in the analysis of many kinds of experiments, and since it is a comparatively new method the possibilities of its application have not yet been exhausted.

Application of Analysis of Variance. As an illustration of the application of the analysis of variance to a series of plot yields we will use the data in Table 90, page 343. The data in this table are from a uniformity test and it is assumed that there were five varieties replicated five times.

Table 90

Application of the Analysis of Variance to a Series of Plot Yields

Varieties	Blocks					Variety Total	Mean
	1	2	3	4	5		
A	51	45	48	48	54	246	49.2
B	47	50	50	54	52	253	50.6
C	48	51	50	51	55	255	51.0
D	52	47	49	55	56	259	51.8
E	48	47	49	50	57	251	50.2
Block Total	246	240	246	258	274	Grand Total	General Mean
Mean	49.2	48.0	49.2	51.6	54.8	1264	50.56

Calculation of Sum of Squares for Variation Due to 'Between Varieties'

Variety Mean—General Mean	D	D^2
49.2 − 50.56	-1.36	1.8496
50.6 − 50.56	.04	.0016
51.0 − 50.56	.44	.1936
51.8 − 50.56	1.24	1.5376
50.2 − 50.56	- .36	.1296
	Sum	3.7120
		×5
		18.5600

Calculation of Sum of Squares for Variation Due to 'Between Blocks'

Block Mean—General Mean	D	D^2
49.2 − 50.56	-1.36	1.8496
48.0 − 50.56	-2.56	6.5536
49.2 − 50.56	-1.36	1.8496
51.6 − 50.56	1.04	1.0816
54.8 − 50.56	4.24	17.9776
	Sum	29.3120
		×5
		146.5600

Values for the Analysis of Variance

Variation Due to Difference	Degrees of Freedom	Sum of Squares	Mean Squares or Variance
Between Varieties	4	18.5600	4.6400
Within Varieties	20	221.6000	11.0800
Between Blocks	4	146.5600	36.6400
Within Blocks	20	93.6000	4.6800
Error	16	75.0400	4.6900
Total	24	240.1600	10.0067

As a guide for similar work a uniform number of decimals have been retained.

The first step is to obtain the total for the varieties and the total for the blocks and the mean for the varieties and the mean for the blocks. The grand total for the entire test is 1264 and the general mean is 50.56. For the analysis of variance we obtain the sum of the squares of the deviations of each individual yield, or plot, from the general mean. For example, for the first yield for variety A we would have the difference between 51 and the general mean, 50.56. The deviations of all of the yields would be obtained, squared, and summed. It is often more convenient to obtain this value by squaring the individual plot yields and then making a correction for the fact that we are not working from the true mean, following steps similar to those used for determining the mean and the standard deviation from an assumed mean value. This method is the one that has been used to obtain the sum of the squares of the deviations of the individual plot yields from the general mean.

Summing the squares of the individual plot yields, $(51)^2+(47)^2+(48)^2+(52),^2$ and so on, we have a total of 64148. This amount must be corrected to give the value that would result if we were working from the true mean. There are several ways of determining this correction. We may square the grand total and divide this by the number of individual plots or square the general mean and multiply by the number of plots. The correction may also be obtained by multiplying the grand total by the general mean. When the total and the number of observations are such that the mean cannot be determined exactly, that is when it is necessary to drop decimals in order to read the mean to a convenient number, then it is more accurate to obtain the correction by squaring the grand total and dividing by the number of observations. This method reduces to a minimum the effect of the decimals used or dropped. It should be remembered that this correction value must be carried to a number of decimal places since it will be used in several calculations and it is important that the net sum of the squares in all cases should be accurate to several decimals. No definite rule need be set, but a laboratory rule should be established so that all calculations will be uniform.

The correction for this problem, obtained by squaring the grand total, 1264, and dividing by the number of individual plots, 25, is 63907.8400. Subtracting this from the total sum of the squares of the individual plot yields, 64148, we have 240.1600. This is called the sum of the squares for total, and appears in the last line of the values for the analysis of variance in Table 90. Since there are 25 plots in all there are 24 degrees of freedom.

The next step is to determine the sums of the squares for the difference between varieties and the difference between blocks. We obtain the sum of the squares for the difference between varieties by taking the difference between the mean yield of each variety and the general mean. For example, the average yield of variety A is 49.2, and the difference between this and the general mean, 50.56, is obtained and squared. Handling the other values in the same way we have 3.7120 for the sum of the squares for the difference between varieties. Since each mean has been obtained from the yields of 5 replications the sum of the squares, 3.7120, is multiplied by 5, giving 18.5600. Since these deviations have been obtained from the true mean it is not necessary to introduce the general correction. For the 5 varieties there are 4 degrees of freedom, and this value, 18.5600, with the degrees of freedom, is recorded in the first row of the values for the analysis of variance in Table 90.

It is also possible to obtain this value by squaring the total yield for each variety, summing these squares, and dividing by 5, since the total yield is made up of 5 plots. The result is 63926.4000, and subtracting from this the correction, 63907.8400, we have as the remainder 18.5600. The total yields may be used for the other calculations in the same way.

The sum of the squares for the difference between blocks is obtained by a similar process. This value, 29.3120, is multiplied by 5, since there are 5 varieties, or 5 plots, in each block. The product, 146.5600, is recorded in the third row of the values for the analysis of variance in Table 90, and since there are 5 blocks there are 4 degrees of freedom.

The sum of the two values, 18.5600 + 146.5600, is obtained and subtracted from the total sum of the squares, 240.1600, giving 75.0400. This is the sum of the squares due to random variation, or due to error. Since this was obtained by summing the two values and subtracting from the total, the degrees of freedom are obtained in the same way by summing the degrees of freedom for the difference between varieties and the difference between blocks and subtracting this sum from the degrees of freedom for the total. Since there are 4 degrees of freedom in each case their sum is 8, and subtracting this from the total degrees of freedom, 24, we have 16 as the degrees of freedom for error.

The analysis of this experiment may be carried still further and the sum of the squares for the variation due to the difference within varieties and the sum of the squares for the variation due to the difference within blocks determined. For the variation due to the difference within varieties we obtain the deviations of the variety means from the individual plot yields for the particular variety. These deviations are squared and summed, giving 221.60, which is recorded in the second line of the values for the analysis of variance in Table 90. For the variation due to the difference within blocks it is necessary to have the deviations between each individual plot yield and the mean of the block in which the plot appears. These deviations are squared and summed, giving 93.60, which appears in the fourth line of the values for the analysis of variance in Table 90. The details of the steps in obtaining these values are given on page 347.

In each case there are 20 degrees of freedom, determined as follows. For the varieties there are 5 plots of each variety, and therefore there are 4 degrees of freedom for each variety. Since there are 5 varieties we have 4 × 5, or 20 degrees of freedom. For the blocks there are 5 varieties in each block, and therefore there are 4 degrees of freedom for each block. As there are 5 blocks we have 4 × 5, or 20 degrees of freedom.

CALCULATION OF SUM OF SQUARES FOR VARIATION DUE TO 'WITHIN VARIETIES'

VARIETY YIELD		MEAN OF VARIETIES	D	D^2
51	−	49.2	1.8	3.24
45	−	49.2	−4.2	17.64
48	−	49.2	−1.2	1.44
48	−	49.2	−1.2	1.44
54	−	49.2	4.8	23.04
47	−	50.6	−3.6	12.96
50	−	50.6	− .6	.36
50	−	50.6	− .6	.36
54	−	50.6	3.4	11.56
52	−	50.6	1.4	1.96
48	−	51.0	−3.0	9.00
51	−	51.0	0.0	0.00
50	−	51.0	−1.0	1.00
51	−	51.0	0.0	0.00
55	−	51.0	4.0	16.00
52	−	51.8	.2	.04
47	−	51.8	−4.8	23.04
49	−	51.8	−2.8	7.84
55	−	51.8	3.2	10.24
56	−	51.8	4.2	17.64
48	−	50.2	−2.2	4.84
47	−	50.2	−3.2	10.24
49	−	50.2	−1.2	1.44
50	−	50.2	− .2	.04
57	−	50.2	6.8	46.24
			SUM	221.60

CALCULATION OF SUM OF SQUARES FOR VARIATION DUE TO 'WITHIN BLOCKS'

BLOCK YIELD		MEAN OF BLOCKS	D	D^2
51	−	49.2	1.8	3.24
47	−	49.2	−2.2	4.84
48	−	49.2	−1.2	1.44
52	−	49.2	2.8	7.84
48	−	49.2	−1.2	1.44
45	−	48.0	−3.0	9.00
50	−	48.0	2.0	4.00
51	−	48.0	3.0	9.00
47	−	48.0	−1.0	1.00
47	−	48.0	−1.0	1.00
48	−	49.2	−1.2	1.44
50	−	49.2	.8	.64
50	−	49.2	.8	.64
49	−	49.2	− .2	.04
49	−	49.2	− .2	.04
48	−	51.6	−3.6	12.96
54	−	51.6	2.4	5.76
51	−	51.6	− .6	.36
55	−	51.6	3.4	11.56
50	−	51.6	−1.6	2.56
54	−	54.8	− .8	.64
52	−	54.8	−2.8	7.84
55	−	54.8	.2	.04
56	−	54.8	1.2	1.44
57	−	54.8	2.2	4.84
			SUM	93.60

It may be of interest to show how these additonal sums of squares may be used in determining the sum of the squares due to error. This may be done by taking the difference between the variation due to the difference within varieties and the variation due to the difference between blocks. We have

VARIATION DUE TO DIFFERENCE	DEGREES OF FREEDOM	SUM OF SQUARES
Within Varieties	20	221.6000
Between Blocks	4	146.5600
ERROR	16	75.0400

We may also obtain this same variation due to error by taking the difference between the variation due to the difference within blocks and the variation due to the difference between varieties, giving

VARIATION DUE TO DIFFERENCE	DEGREES OF FREEDOM	SUM OF SQUARES
Within Blocks	20	93.6000
Between Varieties	4	18.5600
ERROR	16	75.0400

We may now complete the analysis of the experiment. We obtain the mean squares by dividing the sums of the squares by the appropriate degrees of freedom. Dividing the sum of the squares for error, 75.0400, by 16 degrees of freedom, we have 4.6900 as the mean square. That is, after eliminating the effect of the differences between the varieties and the differences between blocks we obtain the variance for error, which is used in the interpretation of the results. Extracting the square root of this variance for error, 4.6900, we have 2.17, which is the standard error of one plot in this experiment. Since we have 5 plots of each variety the standard error of the mean of 5 plots is $2.17/\sqrt{5}$, or .97.

For the general analysis of problems of variance, the variance values for the comparisons studied are compared with the variance values for error. In order to determine whether there is any difference between the variance due to the difference between varieties and the variance due to error, we compare the variance for the difference between varieties with that for error. The mean square for the difference between varieties, 4.6400, is slightly less than that due to error, 4.6900, and this indicates that there is no significant difference between the varieties. This is to be expected since these yields are from a uniformity test, or from a number of plots all sown to the same variety, distributed at random in the field.

If there had been a difference, then we would compare the variances by determining the ratio of the variance due to the difference between varieties to the variance for error. If this ratio were such that it indicated considerable difference between the variance due to the difference between varieties and that due to error, we would conclude that there is a significant difference between the varieties.

A more definite method for comparing the variances has been given by Fisher. This consists of finding the value of $\frac{1}{2}\ log_e$ for each variance and subtracting the value of $\frac{1}{2}\ log_e$ for the variance for error, giving a value designated z for each variance, or we may divide the variance values by the variance due to error and then determine $\frac{1}{2}\ log_e$. The values for $\frac{1}{2}\ log_e$ may be obtained from tables giving the natural logarithms of numbers and taking one-half of the logarithm as tabled. If tables of natural logarithms are not available the common logarithm may be taken and converted to $\frac{1}{2}\ log_e$ by multiplying by the constant 1.1512925. For most calculations it is sufficiently accurate to read this constant to five decimals.

For the resulting z values Fisher has prepared two tables showing how large a difference may be expected due to chance variation, considering the appropriate degrees of freedom. One table is for a 5 per cent level of significance, which is taken to mean that for certain selected degrees of freedom the chances are 1 out of 20 of obtaining as high a z value as tabled due to chance variation alone. The other table is for a 1 per cent level of significance, which indicates how large a value of z may be expected for certain selected degrees of freedom once in a hundred times due to chance variation. In these tables two values for degrees of freedom are given, n_1 and n_2, and for interpreting results between two comparisons n_1 is taken as the number of degrees of freedom corresponding to the item with the larger mean square or variance.

Before leaving the subject of the interpretation of results from the analysis of variance on the basis of Fisher's z values, it may be worth while to give here the method for determining the standard error of z. It is useful as an added guide and for cases where the degrees of freedom lie outside the values tabled. When n_1 and n_2 are large, or when they are of moderate size and nearly equal, the standard error of z is equal to

$$\frac{1}{2}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

When z is larger than twice this value we can say that it lies above the 5 per cent level of significance.

Another table for the analysis of variance, based on Fisher's values, has been prepared by Snedecor, and appears as Table X in the Appendix. For many investigators this table is more convenient since it does not require the use of logarithms. For interpreting results by this table all that is necessary is to determine a value which Snedecor has designated F, and which is the ratio of the larger mean square to the smaller mean square for the variables being compared. The table gives the values of F that may be expected due to chance variation for various degrees of freedom. The degrees of freedom for the variable with the greater mean square are indicated above the columns, and the degrees of freedom for the variable with the smaller mean square are indicated at the left of the rows. When problems are being studied in which the degrees of freedom do not agree with those given in the table, it is usually sufficient to read the values for the degrees of freedom nearest to those of the problem being studied.

It will be observed that there are two values appearing in the table for each comparison, one in light-face type and one in dark-face type. The light-face type indicates the value of F that may be expected due to chance variation alone once out of 20 trials. The dark-face type indicates the value of F that may occur due to chance alone once out of 100 trials. When a value obtained for F for certain degrees of freedom is below the value in light-face type we conclude the result is not significant. If the value obtained for F lies between the number in light-face type and the number in dark-face type, we conclude that the result is significant. If the value of F is larger than the number in dark-face type, we may conclude that the result is highly significant.

An application of this method of interpretation may now be made with the data from Table 90. As already stated, there is really no difference between the variance for 'between varieties' and the variance for error. Merely to illustrate the use of the values of F from Table X we will compare the difference 'within varieties' and the variance for error, although they are not independent. The variance for 'within varieties' is 11.0800 and the variance for error is 4.6900. Therefore $F = 11.0800/4.6900$, or 2.36.

We have 20 degrees of freedom for 'within varieties,' the greater mean square, and 16 degrees of freedom for error, the smaller mean square. Referring to Table X we do not find a column for 20 degrees of freedom and therefore we use 24, the nearest number of degrees of freedom. Reading in this column opposite 16 degrees of freedom we find 2.24 in light-face type and 3.18 in dark-face type. Since the value for F, 2.36, is a little higher than the value in light-face type, 2.24, we conclude that the result is significant, but not highly significant.

We may now refer to Table 90 to see what has been accomplished in the analysis of variance, by separating the total variation into its several component parts. If this had not been done and we used only the total deviation to determine the standard error of the plots we would find from the variance for the total deviation that the standard error is $\sqrt{10.0067}$, or 3.16. It is noted that the mean square obtained from the variation due to the difference between blocks is by far the largest mean square. This indicates a considerable variation from block to block, and when this is eliminated from the total variation it has the effect of greatly reducing the error of the experiment, so that when we take away from the total variation the variation obtained from the difference between blocks and the difference between varieties we reduce the mean square from 10.0067 to 4.6900, and the standard error is reduced from 3.16 to 2.17. Thus we have a lower value for the standard error with which to compare the results of the experiment, and this is due to the fact that we have eliminated certain known effects. One of the important features of the analysis of variance is to eliminate the effects of certain known factors so that a more definite and a more direct comparison may be made between the results being studied.

Before leaving this problem it may be of interest to determine the expected yields for the several plots in Table 90. For example, for the first plot for variety A we may obtain the expected yield for this plot by considering two factors. One is the mean of variety A obtained from all of the plots of this variety, and the other is the comparison of the mean of the block in which this plot falls with the mean of all of the plots in the field. Considering these two factors for each variety we obtain a correction for each plot in the test.

For the first block we will call the correction x_1, and for the other four blocks the corrections will be x_2, x_3, x_4, and x_5. The correction for the first block is obtained by subtracting the general mean from the mean yield of the block. Thus for block 1

$$x_1 = 49.2 - 50.56 = -1.36$$

This correction shows that the first block or replication yields less than the mean of the whole field, and the individual plots must be corrected by this factor. When the correction is minus, the amount of the correction is subtracted from the mean yield of each variety, giving the corrected yield for the individual plots in the block. When the correction is positive it is added. The other corrections are as follows:

For block 2 $x_2 = 48.0 - 50.56 = -2.56$
For block 3 $x_3 = 49.2 - 50.56 = -1.36$
For block 4 $x_4 = 51.6 - 50.56 = 1.04$
For block 5 $x_5 = 54.8 - 50.56 = 4.24$

With these factors the corrected yield for each plot is obtained by applying the correction to the mean yield of the variety. The signs must be observed in making these corrections. The difference between the actual yields and the calculated yields are then obtained, squared, and summed. The details of the process, for the first block, are as follows:

VARIETY	MEAN	x_1	CORRECTED YIELD OF PLOT	OBSERVED YIELD OF PLOT	D	D^2
A	49.2	-1.36	47.84	51	3.16	9.9856
B	50.6	-1.33	49.24	47	-2.24	5.0176
C	51.0	-1.36	49.64	48	-1.64	2.6896
D	51.8	-1.36	50.44	52	1.56	2.4336
E	50.2	-1.36	48.84	48	- .84	.7056

For the other blocks the appropriate corrections, x_2, x_3, x_4, and x_5, are applied to the variety means, observing the signs, and the steps completed as above. The total sum of the squared deviations is

75.0400, which is the same value as was found for error from the analysis of variance. That is, it is the amount of variation remaining, which we may call random variation, after the variation due to the difference between varieties and to the difference between blocks has been eliminated.

This gives a clearer idea of what is being accomplished through the analysis of variance. That is, by eliminating the effect of the difference 'between blocks' and the difference 'between varieties' we obtain a net residue, which is the amount due to random variation or error. This amount may be obtained by the method of the analysis of variance, as already indicated, or by determining the expected or calculated yields, taking the differences between the observed and calculated yields and obtaining the random variation from these differences.

TABLE 91

APPLICATION OF THE ANALYSIS OF VARIANCE IN A COMPARISON OF SIX VARIETIES OF OATS

VARIETIES	YEARS					VARIETY TOTAL	MEAN
	1918	1919	1920	1921	1922		
A	66	52	57	64	61	300	60.0
B	68	52	48	55	53	276	55.2
C	61	48	51	49	56	265	53.0
D	53	46	50	52	43	244	48.8
E	53	47	52	52	47	251	50.2
F	58	41	44	43	48	234	46.8
YEAR TOTAL	359	286	302	315	308	1570	
MEAN	59.8	47.7	50.3	52.5	51.3		

VALUES FOR THE ANALYSIS OF VARIANCE

VARIATION DUE TO DIFFERENCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE	*F* VALUE FOR COMPARISON
Between Varieties	5	575.467	115.0934	8.69
Between Years	4	498.334	124.5835	
Error	20	261.866	13.2433	
TOTAL	29	1338.667		

We may now apply the analysis of variance to a problem where six varieties of oats have been compared for five different years. These data are given in Table 91, page 353.

Following the methods of analysis already given we determine the totals and means for the varieties and years, and then determine the sums of the squares for total and due to the difference between varieties and to the difference between years. Since the steps for obtaining these values have already been explained in detail, the results only are recorded in Table 91.

There are 30 plots in the experiment and therefore there are 29 degrees of freedom for total, and since there are 6 varieties and 5 years we have 5 degrees of freedom for varieties and 4 for years. Summing the sums of the squares and the degrees of freedom for the variation due to the difference between varieties and to the difference between years, and subtracting these from the sum of the squares and the degrees of freedom for total, we have 264.866 remaining as the sum of the squares for error and 20 for the degrees of freedom. The sums of the squares for the difference within varieties and the difference within years may also be obtained, but since we are interested only in learning whether there is any difference between the varieties for the different years it is not necessary to obtain the additional values for this analysis.

Dividing the sums of the squares by the appropriate degrees of freedom we obtain the mean squares, as recorded in Table 91. The mean square for error is 13.2433, and we may now use this to determine the value of F for the comparison of the variance for the difference between varieties and the variance for error. Dividing the mean square for the difference between varieties by the mean square for error, we find F to be 8.69. The larger mean square has 5 degrees of freedom and there are 20 degrees of freedom for error, and from Table X in column 5 opposite 20 we find the two values for F as tabled are 2.71 and 4.10. Since the value for F from the problem is much higher than these, we conclude that there is a very significant difference between the varieties.

Since we have learned that there is a significant difference between varieties as shown by this method of analysis, we will proceed to compare the varieties. It may be stated that while in some instances there may seem to be a considerable difference between say two varieties in a test, unless the whole test leads to significance, as in the present example, it is better not to draw any definite conclusions.

The varieties may be compared by using either the totals or the means. Considering first the totals, we have

VARIETY	TOTAL YIELD
A	300
B	276
C	265
D	244
E	251
F	234

The mean of these totals is 261.6667. Considering this as 100 per cent and expressing each variety as a percentage of the mean, we have

VARIETY	PERCENTAGE OF MEAN
A	114.65
B	105.48
C	101.27
D	93.25
E	95.92
F	89.43

The standard error of a single plot is obtained by taking the square root of the variance due to error. In this case we have

$$\sqrt{13.2433}=3.64$$

as the standard error. Expressing this as a percentage of the mean we have

$$3.64/261.6667=1.39$$

The standard error of a total of 5 plots is $1.39\sqrt{5}$, or 3.11. This may also be obtained by multiplying the variance of a single plot by

5 and extracting the square root, or $\sqrt{13.2433 \times 5}$. This gives a value of 8.14, and expressed as a percentage of the mean we have

$$8.14/261.6637 = 3.11 \text{ per cent}$$

The standard error of the difference is obtained from $3.11\sqrt{2}$, giving 4.40. A difference exceeding twice this value, or 8.80, is considered significant. Using this standard we find variety A is significantly better than variety B, and other comparisons may be made. This illustrates how an experiment may be analyzed and the results interpreted by using the methods of analysis of variance.

The variety means may be used for the interpretation of the results by considering the standard error of the mean. The means of the varieties for the five years of the test are as follows:

VARIETY	MEAN YIELD
A	60.0
B	55.2
C	53.0
D	48.8
E	50.2
F	46.8

The standard error of the mean is

$$3.64/\sqrt{5} = 1.63$$

The standard error of the difference is $1.63\sqrt{2}$, or 2.31, and twice this value is 4.62.

In order that any difference between the means of two varieties may be considered significant it is necessary that the means differ by more than this amount. Variety A gives a mean yield of 4.8 more than variety B, and since this is greater than the measure of significance we conclude that variety A is significantly better than variety B, and it is also better than any of the other varieties in the comparison. Other comparisons may be made, using the same standard to denote significance.

TABLE 92

ANALYSIS OF VARIANCE APPLIED TO A COMPARISON OF THE YIELDS FOR THIRTEEN VARIETIES OF OATS

VARIETIES	BLOCKS										VARIETY TOTAL	MEAN
	1	2	3	4	5	6	7	8	9	10		
A	42	35	41	46	32	36	33	30	22	13	330	33.0
B	41	35	40	43	35	37	29	23	22	20	325	32.5
C	42	34	35	47	37	33	32	31	23	20	334	33.4
D	47	33	36	48	35	41	35	29	27	13	344	34.4
E	45	37	41	44	39	34	30	24	24	17	335	33.5
F	43	43	36	42	37	35	42	33	32	21	364	36.4
G	50	38	34	47	43	37	39	33	28	21	370	37.0
H	45	44	36	56	45	54	39	30	32	23	404	40.4
I	57	47	49	53	45	45	44	33	30	25	428	42.8
J	51	46	58	54	46	43	44	30	34	21	427	42.7
K	44	42	45	48	37	45	48	31	27	19	386	38.6
L	45	44	52	43	39	36	36	37	35	21	388	38.8
M	42	41	31	40	34	46	38	27	27	18	347	34.7
BLOCK TOTAL	594	522	534	611	504	522	489	391	363	252	4782	GENERAL MEAN 36.7846

VALUES FOR THE ANALYSIS OF VARIANCE

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE	*F* VALUE FOR COMPARISON
Varieties	12	1567.5692	130.6308	8.40
Blocks	9	8524.5846	947.1761	
Error	108	1679.8154	15.5538	
TOTAL	129	11771.9692		

As suggested earlier, we should be careful in drawing conclusions from experiments similar to this if the analysis of the whole experiment does not show significance for the varieties or for the treatments which are being analyzed. That is, it is possible that the whole experiment may show no significance, and yet one might expect to take an extremely poor result, such as a low yield, and compare it with an extremely high yield and obtain significance on the basis of the standard error. This should not be done unless the result of the entire experiment shows significance.

Another illustration giving the results of the analysis of a larger experiment may be shown. The data are given in Table 92, page 357, and are the results obtained from comparing 13 varieties distributed at random, with 10 blocks of each variety.

Since the methods for obtaining the several values for the analysis of variance are simple and straightforward it is not necessary to repeat them in detail here. The results are included in the table, showing the sums of the squares for total, due to the variation between varieties and to the variation between blocks, and the remainder due to error.

The mean square for the difference between varieties is 130.6308 and the mean square for error is 15.5538, and F is 8.40. Referring to Table X we find that for 12 and 108 degrees of freedom the nearest values tabled are 1.85 and 2.37. Since the value for F obtained from the problem, 8.40, is much higher than these, we conclude that there is a very significant difference between the yields of the varieties, and we may proceed to compare the varieties.

Obtaining the square root of the variance for error we find that the standard error of a single plot is 3.94, and the standard error of the mean of 10 plots is

$$3.94/\sqrt{10}=1.25$$

For the standard error of the difference we have $1.25\sqrt{2}$, or 1.77, and twice this value is 3.54.

Any of the varieties whose mean yields differ by this amount may be said to be significantly different so far as their yielding capacity is concerned. Thus variety G is better than varieties A and B, but it is poorer than varieties I and J. There is no significant difference between varieties H, I, and J. Other comparisons may also be made.

As explained in the previous problem, the varieties may be compared by expressing the mean yields as a percentage of the general mean and obtaining the standard error as a percentage value also. Taking the general mean, 36.7846, as 100 per cent, the means for the 13 varieties are as follows:

VARIETY	MEAN IN PER CENT
A	89.71
B	88.35
C	90.80
D	93.52
E	91.07
F	98.95
G	100.59
H	109.83
I	116.35
J	116.08
K	104.94
L	105.48
M	94.33

Expressing the standard error of the mean in per cent we have

$$1.25/36.7846 = 3.40 \text{ per cent}$$

Obtaining the standard error of the difference and taking twice this value we have 9.62 per cent as the standard for comparing the different varieties. Any two varieties may be compared, and if they do not show a percentage difference of 9.62 or more they would not be considered significantly different.

The methods for the analysis of variance that have just been discussed are those appropriate for the analysis of results that have been obtained from plots arranged by random with one restriction, and that is that only one plot of each treatment or variety was allowed in each replication or block. We will now consider a further restriction of the random arrangement of plots.

Suppose that five varieties are to be tested in five replicated plots in a field marked out as in Figure 33.

	Columns				
Rows	1	2	3	4	5
1	D	A	C	B	E
2	B	E	D	C	A
3	C	B	A	E	D
4	E	D	B	A	C
5	A	C	E	D	B

FIG. 33. Proposed arrangement of Latin square for five varieties.

In the first row the varieties may be arranged at random, having only one plot of each variety in row 1. This restriction is the same as that just considered, but in addition the arrangement of the varieties is further restricted so that only one plot of each variety may appear in any one column. Thus we have a double restriction, and the varieties may be distributed as indicated. Such an arrangement is known as the Latin square, and has been given much prominence by Fisher in his work on the analysis of variance.

There are various forms and arrangements of Latin squares, but at all times the considerations are that the arrangement of the treatments or varieties in the rows and in the columns shall be restricted so that only one plot of each treatment or variety shall appear in any one row or in any one column. In a Latin square the general form is to have as many replications as there are treatments or varieties. While the arrangement of the Latin square suggests a field where it is possible to lay out plots in the form of a square, it may be stated that if it is necessary to have the plots in strips arranged side by side, so long as the same restrictions have been followed the method for the analysis of the Latin square may still be used. Various plans may be followed in developing the forms of the Latin square, as is true for random arrangement.

TABLE 93

ARRANGEMENT AND ANALYSIS OF A LATIN SQUARE

	COLUMNS					TOTAL	MEAN
ROWS	D 37	A 38	C 38	B 44	E 38	195	39.0
	B 48	E 40	D 36	C 32	A 35	191	38.2
	C 27	B 32	A 32	E 30	D 26	147	29.4
	E 28	D 37	B 43	A 38	C 41	187	37.4
	A 34	C 30	E 27	D 30	B 41	162	32.4
TOTAL	174	177	176	174	181	GRAND TOTAL 882	GENERAL MEAN
MEAN	34.8	35.4	35.2	34.8	36.2		35.28

RESULTS FOR VARIETIES

VARIETY	A	B	C	D	E
TOTAL	177	208	168	166	163
MEAN	35.4	41.6	33.6	33.2	32.6

ROWS		COLUMNS		VARIETIES	
N	N^2	N	N^2	N	N^2
195	38025	174	30276	177	31329
191	36481	177	31329	208	43264
147	21609	176	30976	168	28224
187	34969	174	30276	166	27556
162	26244	181	32761	163	26569
SUM	157328	SUM	155618	SUM	156942
	5)157328		5)155618		5)156942
	31465.60		31123.60		31388.40
	31116.96		31116.96		31116.96
	348.64		6.64		271.44

TABLE 93—*Continued*

VALUES FOR THE ANALYSIS OF VARIANCE

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE	*F* VALUE FOR COMPARISON
Rows	4	348.64	87.1600	
Columns	4	6.64	1.6600	
Varieties	4	271.44	67.8600	4.32
Error	12	188.32	15.6933	
TOTAL	24	815.04		

The method of the analysis of the Latin square is similar to the methods that have already been given, with certain modifications. For the analysis of a Latin square we eliminate the effect of rows and columns from the total variation, and also eliminate the effect of treatments or varieties, leaving to residue the sum of the squares due to random variation, or error.

An illustration of the arrangement and analysis of a Latin square is given in Table 93, pages 361, 362. Here there are five varieties distributed in the manner just described. The totals and means for the rows and the columns are obtained, and the totals and means for each of the varieties are also obtained, as shown in Table 93. Thus for variety A we sum the yields for each plot of A, and for variety A from row 1 we have 38, from row 2 we have 35, and so on.

The total sum of the squares is obtained in the usual manner by squaring the individual plot yields, summing, and subtracting the correction, which has been obtained in this case by multiplying the grand total by the general mean. Next, the sums of the squares for rows, columns, and varieties are obtained, as illustrated in the table. Since each of these three totals is made up of the results of 5 separate plots, the sum of the squares is divided by 5, and the correction is subtracted from the quotient. The sums of the squares for total and for rows, columns, and varieties, with their appropriate degrees of freedom, are recorded in Table 93 under the values for the analysis of variance. The sums for the rows, columns, and varieties are added together and subtracted from the total sum of the squares, leaving 188.32 as the sum of the squares due to random variation, or error. Obtaining the degrees of freedom for error by summing the degrees of freedom for rows, columns, and varieties, and subtracting this sum from the total degrees of freedom, we have 12 degrees of freedom left for error. Dividing the several sums of squares by the appropriate degrees of freedom we have the values for the variances. It is noted that the variance for columns is very low, and this is apparent from the total yields of the columns, where the yields are very similar.

In order to learn whether there is any significant difference between varieties, we may now determine the F value. This is done by dividing the variance for varieties, 67.8600, by the variance for error, 15.6933, giving 4.32. Referring to Table X we find that this value for F lies between the two values tabled for the degrees of freedom concerned. Thus we may conclude that there is a significant difference between the varieties, but that the difference is not highly significant.

Since the analysis shows that there is a significant difference between varieties, we may now proceed to analyze the results. This may be done by using the totals and expressing the total yields of the varieties as a percentage of the mean of the total yield of all of the varieties. The standard error will also be expressed in per cent.

VARIETY	PERCENTAGE OF MEAN
A	100.3
B	117.9
C	95.2
D	94.1
E	92.4

The standard error of a single plot is the square root of the variance for error, or $\sqrt{15.6933} = 3.96$. In the total yields there are 5 plots involved, so the standard error of the total of 5 plots will be the standard error of a single plot multiplied by $\sqrt{5}$. We have

$$3.96\sqrt{5} = 8.85$$

The mean of the total yields of all of the plots is 176.4, and expressing the standard error of the total yield in per cent we have

$$8.85/176.4=5.02 \text{ per cent}$$

The standard error of the difference between two means is therefore 7.10, and twice this value is 14.20.

Any varieties that show a greater difference than this may be considered as significantly different. Variety B is 17.6 per cent better than variety A and therefore it may be considered as being significantly better. Variety B may also be compared with the other varieties, and it is found to be significantly better than the others. There is no significant difference between varieties A, C, D, and E.

This illustrates the method of analysis for a Latin square, and shows the possibility of reducing the error by being able to eliminate the effect of both rows and columns. This same method of analysis is useful for the interpretation of results from a fertilizer test where the plots have been arranged in a Latin square.

Eliminating the Effect of Missing Plots. From the foregoing illustrations it is evident that the analysis is simple enough providing results are obtained from each plot or from each item in the experiment. There are unfortunately many factors which operate to affect the result of one plot in such a way that it is not truly representative of the experiment as a whole. Occasionally a plot is entirely destroyed because of crowding a road or path out on the plot, or due to injury by cattle, or in many other ways.

In case of such accidents it is important to have some way of estimating the yield of this missing plot. Various suggestions have been made, but the best method available at present seems to be that suggested by Allan and Wishart, and which has been further elaborated by Yates. Without giving the arguments leading to the formula for the calculation of a missing plot yield, we will proceed to apply the formula. The correction is made on the basis of the variation occurring within the plots receiving the same kind of treatment, and within the column in which the missing plot occurs. We will use the data in Table 92 and assume that one plot is missing. This is the plot in the second block for variety C. In the table the yield of this plot is given as 34, but we will assume that it is missing and calculate the yield.

The argument presented by Allan and Wishart leads to the following formula for the calculation of the yield of a missing plot for randomized blocks:

$$x=\frac{pP+qQ-T}{(p-1)(q-1)}$$

in which

x = the yield of the missing plot
p = the number of varieties
q = the number of blocks
P = the sum of the yields of all of the plots of the same variety
Q = the sum of the yields of all of the plots in the block in which the missing plot occurs
T = the total yield of all of the plots from which yields are available.

Assuming that this plot is missing, its yield must be deducted from the row and column in which it occurs, and also from the total yield of all of the plots. We have

$$\begin{aligned} P &= 334-34=300 \\ Q &= 522-34=488 \\ T &= 4782-34=4748 \end{aligned}$$

Substituting these values in the formula we have

$$x=\frac{(13\times 300)+(10\times 488)-4748}{12\times 9}=\frac{4032}{108}=37.333$$

Reading the result to a whole number we find that the calculated yield for the plot is 37, while the observed yield is 34. This indicates that the calculated plot yields may always be subject to some variation, and that there is no method that will lead at all times to the exact results for missing plots.

It will be important to see what effect this change in yield has on the complete analysis of the experiment. Substituting 37 for 34 and completing the analysis of variance in the usual way, allowing for one degree of freedom lost in calculating the yield of the missing plot, we have

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE
Varieties	12	1548.0923	129.0077
Blocks	9	8545.4231	949.4915
Error	107	1670.6769	15.6138
TOTAL	128	11764.1923	

The standard error, which is the square root of the mean square for error, is 3.95. It is seen that the standard error has not been influenced to any great extent, as the standard error determined from the original experiment in Table 92 is 3.94.

The application of this formula has been extended by Yates to cases where more than one plot may be missing, and as a result of involved mathematical analysis, which need not be given here, he finds that by the use of the formula above, with certain modifications and obtaining a second approximation, the results agree very well with those obtained from his more elaborate analysis. The simpler method will be given here, and using the data in Table 92 we will assume that four plots are missing. The missing plots are underlined in the table.

The first step is to determine the total yield of all of the plots whose yields are available, and from this the mean yield is calculated by dividing by the number of plots. The total yield of all of the plots in Table 92 is 4782, and considering the four underlined plots as missing this total becomes

$$4782-(34+47+24+37)=4640$$

This is divided by the number of individual plot yields remaining, 126, and the mean yield is 36.83.

The yields of the missing plots are determined from the formula just used, calculating one plot at a time. For each calculation the yield of the other three missing plots is assumed to be equal to the mean yield of all of the plots. Therefore three times the mean yield, or

$$3\times 36.83=110.49$$

is added to the total yield of the 126 plots, giving a total of 4750.49, which is the total that will be substituted in the formula.

The value of P for this problem is obtained by subtracting the yield of the missing plot from the total yield of the variety containing the missing plot, and the value of Q is obtained by subtracting the yield of the missing plot from the total yield of the block in which the missing plot occurs. For the missing plot in block 1 for variety D

$$P = 344 - 47 = 297$$

and

$$Q = 594 - 47 = 547$$

Substituting these values in the formula, remembering that p is the number of varieties and q is the number of blocks, we have

$$x_1 = \frac{(13 \times 297) + (10 \times 547) - 4750.49}{12 \times 9} = \frac{4580.51}{108} = 42.41$$

Obtaining the values for P and Q for the missing plots in the other blocks in a similar manner and substituting these values in the formula, the calculated yields, or first approximations, for the four missing plots in the order of their occurrence are

	First Approximation
x_1	42.41
x_2	37.31
x_3	41.26
x_4	27.43

The sum of these calculated yields is obtained and added to the total, 4640, giving a new total of 4788.41. This new value for the total is used in the formula for obtaining the second approximations.

The same formula is used for calculating the yields of the missing plots in the second approximation, but a change is made in the total value. The values of P and Q are the same as before, but for T we substitute the new total, 4788.41, minus the first approximation yield of that plot for which we are calculating the result. For x_1 we have

$$x_1=\frac{(13\times 297)+(10\times 547)-(4788.41-42.41)}{12\times 9}=\frac{4585.00}{108}=42.45$$

The values for the other missing plots are obtained in a similar way, subtracting from 4788.41 the yield of the missing plot for which the calculation is being made. This is done for each missing plot, and the values obtained for the second approximation, together with those for the first approximation are

	FIRST APPROXIMATION	SECOND APPROXIMATION
x_1	42.41	42.45
x_2	37.31	37.30
x_3	41.26	41.29
x_4	27.43	27.33

In this example the differences between the first and second approximations are very small. In some cases the differences will be larger and the second approximation will be more nearly the true value. In order to compare the effect that these calculated yields have on the standard error for this experiment, the calculated yields are read to whole numbers and substituted in place of the observed yields, and the analysis of variance completed, correcting the degrees of freedom for the missing plots. The standard error has been calculated and is 3.96, which is slightly larger than the standard error for the whole experiment, 3.94.

It is well to point out that we cannot expect the calculated yields to agree so closely with the observed yields in all cases, which suggests that the calculation of missing plot yields can at best be only an approximation. If only a few plots are missing from a comparatively large number, then it would be desirable to calculate the expected yields and complete the analysis of variance. If a larger number of plots in proportion to the total number are missing, then less reliance can be had in the calculated yields.

As stated, the formula that has been used is that for randomized blocks. For a Latin square the formula is modified, and is as follows:

$$x=\frac{p\,(P_r+P_c+P_t)-2T}{(p-1)\,(p-2)}$$

in which

p = number of treatments or varieties
P_r = known yield of the row containing the missing plot
P_c = known yield of the column containing the missing plot
P_t = known yield of treatment, or variety, containing the missing plot
T = total yield.

While these formulas are to be used for calculating the yield of a missing plot or plots, it is also possible to use them in cases where for some reason it is evident that the yield of a certain plot does not represent the true condition. That is, the yield of a plot may be so high or so low that it indicates that something must have happened to the records, or perhaps the location of the plots in the field was such that while we cannot count the plot as missing yet it was unduly influenced by certain environmental factors so that it seems wise to eliminate it from the comparisons. In such cases careful judgment must be exercised and plots should not be eliminated too freely.

Application of Variance to Regression Analysis. In the study of regression in Chapter VI the standard deviation of the estimate based on regression was determined. We have now learned that this may be referred to also as the standard error of estimate. It is possible to apply the idea of the analysis of variance to regression analysis and the determination of the standard error of estimate, and we have the following general plan for such an analysis.

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE
Regression	1	Total sum of squares times r^2	Sum of squares divided by degrees of freedom
Amount unaccounted for, or residual	N–2	Total sum of squares times $(1-r^2)$*	
TOTAL	N–1	Total sum of the squares measured from the mean of the series	

*Also obtained by subtracting values in first line from Total.

In analyzing the effects of regression in this manner we determine first the total sum of the squares of the deviations of each individual from the mean. This is recorded opposite 'Total' under the column headed 'Sum of Squares.' The degrees of freedom will be one less than the number of individuals in the correlation analysis.

We may now apply this method to the data for the x distribution in Table 36 of Chapter VI. The total sum of the squares may be conveniently obtained from the value for ΣfD^2 as usually calculated, by subtracting the proper correction for working from an assumed mean. For the data in Table 36 the value for ΣfD^2 is 540, obtained by working from an assumed mean. The correction is .485. This correction is squared and multiplied by 400 and subtracted from 540, leaving 445.91 as the total sum of the squares. Since there are 400 individuals there are 399 degrees of freedom.

We may now take out of this total sum the amount due to regression. This is obtained by multiplying the total sum of the squares by the square of the correlation coefficient. For the data in Table 36 the correlation coefficient is .217, and squaring this and multiplying by the total sum of the squares we have 20.9975 as the sum of the squares due to regression. This amount is subtracted from the total sum of the squares, leaving 424.9125 as the amount that is unaccounted for by regression. There was 1 degree of freedom used in determining regression, and this leaves 398 degrees of freedom for the variation not accounted for. The complete results are as follows:

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE
Regression	1	20.9975	20.997500
Amount unaccounted for, or residual	398	424.9125	1.067619
TOTAL	399	445.9100	

The mean square for the amount unaccounted for, or residual, is 1.067619, and the square root of this value, or 1.033, is the standard deviation or standard error. This is the standard error of estimate and may be compared with the standard error of estimate as obtained from the usual formula, given in Chapter VI. By this formula

$$S_x = \sigma_x \sqrt{1-r^2}$$

the standard error of estimate was found to be 1.031. The slight difference in the two results is due to the decimals retained in the calculations. This shows that it is possible to determine the standard error of estimate by the application of the method of variance analysis in the manner just described.

It will be of interest to apply this method to the problem in multiple correlation which was discussed in Chapter VIII. In this chapter the simple correlation between two characters, Y and A, was determined and found to be .654, as recorded in Table 58. Using this value for the correlation coefficient and the total sum of the squares for Y we may determine the error of estimate. Referring to Table 58 we obtain the total sum of the squares for Y, 1335.27, and since there are 25 individuals concerned there are 24 degrees of freedom. Multiplying this total sum of the squares by the square of r, .654, we obtain 571.1163 as the amount due to regression. The amount unaccounted for is the difference between this value and the total sum of the squares, or it may be obtained by multiplying the total sum of the squares by $1-r^2$. The values for this analysis are

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE
Regression	1	571.1163	571.116300
Amount unaccounted for, or residual	23	764.1537	33.224074
TOTAL	24	1335.2700	

Obtaining the standard error for the variation unaccounted for we find that it is 5.764, while by the formula for determining the standard error of estimate in Chapter VIII it was found to be 5.763. The standard error (standard deviation) of estimate measured from the mean is obtained by dividing the total sum of the squares by the degrees of freedom and extracting the square root, and it is found to be 7.459, as in Chapter VIII.

In the problem on multiple correlation the analysis was carried further and the effect of two characters, A and B, on Y was determined. The coefficient of multiple correlation in this case was found to be .700. This coefficient of multiple correlation may be used and the standard error of estimate determined in the manner just described. We have

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE
Regression	2	654.2823	327.141150
Amount unaccounted for, or residual	22	680.9877	30.953986
TOTAL	24	1335.2700	

It should be noted that in this case since we have added another variable in determining the coefficient of multiple correlation we also use one degree of freedom, leaving one less degree of freedom for the amount unaccounted for. The standard error of estimate may be obtained by dividing the sum of the squares unaccounted for by the degrees of freedom, and extracting the square root. The standard error is found to be 5.564, the same as was determined by the formula for the standard error of estimate in Chapter VIII.

Continuing for the effect of three variables, we have .866 as the coefficient of multiple correlation. Applying this value as before we have

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE
Regression	3	1001.3937	333.797900
Amount unaccounted for, or residual	21	333.8763	15.898871
TOTAL	24	1335.2700	

One additional degree of freedom is used since another variable has been added, and we have 21 degrees of freedom left for the amount unaccounted for. Dividing the sum of the squares for residual by 21 and extracting the square root, we have 3.987 as the standard error of estimate, which agrees with the value obtained in Chapter VIII. This method of analysis gives another measure of the standard error of estimate and illustrates how the degrees of freedom change with the addition of more variables in the determination of the regression. Thus it is possible to determine the standard error of estimate by either method, with practically the same results.

It may also be pointed out that the significance of r or R may be determined from the values in the analysis of variance. Taking the last example, we may determine the significance of R by obtaining the value of F from 333.797900/15.898871, or 21 00. Referring to Table X, with degrees of freedom of 3 and 21 we find that this value of F is much larger than either of the values tabled, indicating that R .866 is highly significant. Similar comparisons may be made from the other examples. It may also be pointed out that the ratio of the sum of the squares due to regression to the sum of the squares for total is equal to R^2. For example, in the case just cited, $R^2 = 1001.3937/1335.2700$, or .749956, and $R = .866$.

We may also apply the method of variance analysis to determine the linearity of regression, by obtaining the following values.

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE
Linear regression	1	Total sum of squares times r^2	Obtained by dividing sum of squares by degrees of freedom
Deviations from linear regression	$n-2$	Total sum of squares times $(\eta^2 - r^2)$	
Residual within arrays	$N-n$	Total sum of squares times $(1-\eta^2)$	
TOTAL	$N-1$	Total sum of the squares measured from the mean of the series	

The method of determining the effect of linear regression has already been explained, and we need now to add another item, namely the effect of the deviations from linear regression.

The total sum of the squares due to linear regression is obtained by multiplying the total sum of the squares as measured from the mean by r^2. The deviations from linear regression are obtained by determining the difference between the squares of the correlation ratio and the correlation coefficient, and multiplying the total sum of the squares as measured from the mean by this difference. The degrees of freedom for the total are as usual $N-1$, and 1 degree of freedom is used for linear regression. The degrees of freedom for the variation due to the deviations from linearity, or linear regression, are given as $n-2$, n referring to the number of arrays used in the calculation of the correlation ratio. The degrees of freedom left for the sum of the squares for the residual are equal to $N-n$, and they may be arrived at in the usual way by summing the degrees of freedom for linear regression and for the deviations from linear regression and subtracting this sum from the degrees of freedom for the total.

We may now apply this method of analysis to the data in Table 48 of Chapter VII. The correlation coefficient for these data is .865, and η_{yx} is .904. In obtaining this correlation ratio there were 11 arrays used, therefore $n = 11$. The total sum of the squares is obtained in the usual way, giving 442.9167. Multiplying this value by the values for r^2 $(\eta^2 - r^2)$, and $(1-\eta^2)$, as in the plan outlined above, we have

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE
Linear regression	1	331.4013	331.4013
Deviations from linear regression	9	30.5573	3.3953
Residual within arrays	289	80.9581	.2801
TOTAL	299	442.9167	

In order to ascertain whether there is any decided deviation from linearity it is necessary to determine whether the deviations from linear regression are greater than would be expected from chance variation. This may be done by comparing the variance due to deviations from linear regression with the variance from the effect of 'residual within arrays.' If there is no real departure from linearity the two variances will be nearly equal. If there is an appreciable departure from linearity, then the mean variance of row 2 will be greater than the mean variance of row 3. Dividing the sums of the squares by the appropriate degrees of freedom, the mean variance for row 2 is 3 3953 while for row 3 the mean variance is .2801. It is apparent that there is a considerable difference between these two values. We may now divide the variance in row 2 by the variance in row 3 and obtain F, which is 12.12 Referring to Table X we find that we do not have values for the degrees of freedom used here, but for $n_1 = 8$ and $n_2 = 300$ we find an F value of 2.57 for the higher degree of significance. Since the F value for this experiment is much larger than 2.57 it indicates that there is a real deviation from linearity.

This is a more exact method for determining whether there is any important deviation from linearity than the one referred to in Chapter VII, since it is possible to separate the total variance into its several components and allowance is made for the number of arrays.

The data in Table 48 were used to illustrate the fitting of curved regression lines, with the results recorded in Table 50. We may use the method of variance to analyze the results obtained by fitting the curved regression lines. For this analysis we determine the deviations from the mean of the observed values as given in the first column of Table 50. The mean of this series is 2.007. Obtaining the difference between each of the individual observations and the mean, squaring, and summing, we have 25.8669, and since there are 11 items there will be 10 degrees of freedom.

The amount of variation due to straight line regression is 21.7827, which is determined from the formula

$$\frac{N\,(N^2-1)}{12}B^2$$

This is the formula giving the amount of reduction from the total due to linear or first order regression. In this and the following formulas N refers to the number of items, and B, C, and D to the values determined in fitting curved regression lines in Chapter VII. On reviewing the discussion on curved regression the meaning of these formulas will be clear.

The amount of reduction due to the second order regression is obtained from

$$\frac{N\,(N^2-1)\,(N^2-4)}{180}C^2$$

and from this we have the value 3.9674. The amount of reduction due to the third order regression is obtained from

$$\frac{N\,(N^2-1)\,(N^2-4)\,(N^2-9)}{2800}D^2$$

giving .0556.

The values for the analysis of variance are

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE
First order regression	1	21.7827	21.7827
Second order regression	1	3.9674	3.9674
Third order regression	1	.0556	.0556
Residual	7	.0612	.0087
TOTAL	10	25.8669	

Summing the amounts due to the first, second, and third order regressions and subtracting this sum from the total sum of the squares leaves a very small amount, .0612, for the residual, or the amount still unaccounted for in the regression analysis. There is 1 degree of freedom for each order of regression, so 7 degrees of freedom are left for the residual.

It is evident that the mean variances for the first order and second order regressions are considerably higher than the variance for the residual, so they are both significant. We may test the third order regression value by determining F, or $.0556/.0087 = 6.39$, and with degrees of freedom of 1 and 7 it is found to be significant. It is evident that there has been considerable reduction in the mean variance as we proceed from the first to the third order regression. The work may be carried further, but from this result and from the values in the last column of Table 50 it is evident that the observed values are well followed by the curved regression line, and the values predicted from this line will have a small error.

The foregoing examples show the usefulness of the methods for variance analysis. The examples given illustrate some of the applications, and it is possible to extend the methods to other kinds of problems, such as the results from soil experiments, feeding or nutrition experiments, and so on. Since by the application of the methods of variance it is possible to divide the total variation into different components, smaller differences may be measured and a more exact analysis made than can be done by the usual methods of determining variability and probable errors. These latter methods determine the errors for the whole experiment without eliminating the effect due to any one or more known causes.

CHAPTER XIV

ANALYSIS OF VARIANCE—COMPLEX EXPERIMENT

The methods that have been given for the analysis of variance are for simple experiments and the calculations are straightforward. It is often important to extend the methods for the analysis of more complex experiments. As an illustration of the application of analysis of variance in such cases the data in Table 94 are used. These are the yields in bushels per acre of five varieties of wheat, replicated five times on two fields in 1928, 1929, 1930, and 1931. The five varieties are designated A, B, C, D, and E.

TABLE 94

YIELDS IN BUSHELS PER ACRE OF FIVE VARIETIES OF WHEAT, REPLICATED FIVE TIMES ON TWO FIELDS IN 1928, 1929, 1930, AND 1931

FIELD I

1928

BLOCKS	VARIETIES					BLOCK TOTAL
	A	B	C	D	E	
1	23	29	26	20	31	129
2	20	29	21	18	20	108
3	17	19	26	13	17	92
4	24	31	30	18	24	127
5	20	29	22	18	21	110
VARIETY TOTAL	104	137	125	87	113	566

TABLE 94—*Continued*

1929

BLOCKS	VARIETIES					BLOCK TOTAL
	A	B	C	D	E	
1	19	19	23	25	29	115
2	23	22	24	29	24	122
3	30	33	36	25	27	151
4	31	31	30	28	34	154
5	25	26	22	20	24	117
VARIETY TOTAL	128	131	135	127	138	659
1930						
1	21	21	24	22	33	121
2	31	25	46	54	58	214
3	49	56	49	56	62	272
4	39	53	44	38	51	225
5	31	41	62	37	46	217
VARIETY TOTAL	171	196	225	207	250	1049
1931						
1	48	50	42	46	45	231
2	51	51	42	49	46	239
3	38	38	36	45	39	196
4	43	46	40	47	38	214
5	43	36	35	29	36	179
VARIETY TOTAL	223	221	195	216	204	1059

FIELD II

1928

BLOCKS	VARIETIES					BLOCK TOTAL
	A	B	C	D	E	
1	21	27	31	18	28	125
2	27	29	40	15	13	124
3	24	30	32	21	21	128
4	30	33	39	28	21	151
5	20	25	37	30	20	132
VARIETY TOTAL	122	144	179	112	103	660

TABLE 94—*Continued*

1929

BLOCKS	VARIETIES					BLOCK TOTAL
	A	B	C	D	E	
1	30	23	33	19	24	129
2	19	24	25	21	25	114
3	25	28	28	21	20	122
4	19	27	18	18	32	114
5	22	22	24	22	17	107
VARIETY TOTAL	115	124	128	101	118	586
			1930			
1	43	18	53	38	50	202
2	28	9	31	57	53	178
3	37	27	25	49	47	185
4	27	13	34	23	27	124
5	32	20	29	36	37	154
VARIETY TOTAL	167	87	172	203	214	843
			1931			
1	41	48	56	47	43	235
2	36	30	54	46	25	191
3	36	35	31	33	33	168
4	46	32	37	39	39	193
5	37	20	31	46	31	165
VARIETY TOTAL	196	165	209	211	171	952

Before proceeding to the analysis of the complex experiment we will analyze the results from Field I for 1928 in the usual manner. Since there are 25 plots in the test there will be 24 degrees of freedom for total. There are 5 varieties and 5 blocks, and therefore there will be 4 degrees of freedom for each, leaving 16 degrees of freedom attributable to error.

To obtain the sum of the squares for total, each of the 25 yields is squared and summed, and from this sum is subtracted a correction value. This correction value is obtained by squaring the grand total and dividing by the total number of yields. We have

$(566)^2/25$, or 12814.24, for the correction. The sum of the squares of the 25 yields is 13424, and deducting 12814.24 from this we have 609.76 as the sum of the squares for total.

For the sum of the squares for varieties, each variety total is squared and the sum of the squares is divided by the number of blocks contained in each variety total. We have 65548 for the sum of the squares, and dividing by 5, the number of blocks in each variety total, we have 13109.60. From this is subtracted the correction, 12814.24, leaving 295.36 as the sum of the squares for varieties.

The sum of the squares for blocks is obtained similarly. Each block total is squared and summed, giving 64998. This is divided by the number of varieties in each block total, or 5, giving 12999.60. Subtracting from this the correction, 12814.24, the remainder is 185.36, which is the sum of the squares for blocks.

The analysis of variance is

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE
Varieties	4	295.36	73.840
Blocks	4	185.36	46.340
Error	16	129.04	8.065
TOTAL	24	609.76	

The sum of the squares attributed to error is obtained by subtracting the sums of the squares for varieties and blocks from the sum of the squares for total, and the degrees of freedom for error are obtained by a similar subtraction. The mean squares are obtained by dividing the sums of squares by their respective degrees of freedom. It is evident from the difference between the mean square for error and the mean square for varieties that there is a significant difference between the yields of the varieties for this test.

For the complete analysis of the entire experiment it is necessary to determine the variance due to varieties, fields, years, and blocks, and the interaction of these factors, and the variation due to each factor and to the interactions are determined separately. The results are brought together for final analysis in Table 95.

The degrees of freedom are the total number of observations, less 1. For the entire experiment there are 200 observations, and the degrees of freedom are 200 − 1, or 199. Similarly, for varieties, blocks, fields, and years the degrees of freedom will be one less than the number observed. For the first order interaction of these various factors, the degrees of freedom will be the product of the degrees of freedom of the factors concerned. For the second order interactions the degrees of freedom will be the product of the degrees of freedom of the three factors concerned. The degrees of freedom for error are obtained by subtracting from the total degrees of freedom the sum of the several individual degrees of freedom. In the analysis of variance for the data in Table 94 the degrees of freedom are as follows:

VARIATION DUE TO	DEGREES OF FREEDOM
Varieties	4
Fields	1
Years	3
Blocks	4
Interaction of:	
Varieties × fields	4
Varieties × years	12
Fields × years	3
Varieties × fields × years	12
Blocks × years	12
Blocks × fields	4
Blocks × fields × years	12
Error	128
TOTAL	199

For this analysis the sum of the squares and degrees of freedom for the entire experiment are determined first. Each of the individual yields in Table 94 is squared, and the sum of these squares is 228414. From this total it is necessary to deduct a correction value, 203139.38, which is obtained by squaring the sum of the individual yields, or grand total, 6374, and dividing by the total number of observations, 200. The remainder, 25274.62, is the sum of the squares for total and the corresponding degrees of freedom are 200 − 1, or 199. This value is recorded in line 13 of Table 95.

It is necessary to correct the total sum of the squares in the manner just outlined, since what is wanted is the total sum of the squares of the deviations of the individual plot yields from the mean of all the plots. Unless the numbers are too large it is usually more convenient to square the yields directly rather than to take the deviation of each one from the mean and then square the deviation. Since the method that will be followed in this analysis is that of squaring the values themselves, it will be necessary in all cases to correct for the fact that we are not working from the true mean but really from an assumed mean of zero, and the same correction will be applied throughout the analysis of the experiment.

For the variation due to varieties, fields, and years, and their interactions, and for the subsequent analyses, we combine the yields from Table 94 according to these various factors. We have first the total yields for varieties and fields.

TOTAL YIELDS FOR VARIETIES AND FIELDS

FIELDS	VARIETIES					TOTAL
	A	B	C	D	E	
I	626	685	680	637	705	3333
II	600	520	688	627	606	3041
TOTAL	1226	1205	1368	1264	1311	6374

These individual values are obtained by combining the yields of each variety for the period of the test. For example, for variety A on Field I, the yields are 104 + 128 + 171 + 223, or a total of 626 for the four years. For variety A on Field II, the yields are 122 + 115 + 167 + 196, giving a total of 600 for the four years, and so on. The sum of the yields of all varieties on each field gives the field total, and the sum of the yields on each field gives the variety total.

For the variation between varieties, each variety total is squared, and the sum of these squares is divided by the number of blocks contributing to the total yield. Thus we have

$$(1226)^2+(1205)^2+(1368)^2+(1264)^2+(1311)^2=8142942$$

This sum is divided by 40, since each variety total is derived from 5 blocks on 2 fields during 4 years, and the quotient is 203573.55. From this is subtracted the correction obtained for the entire experiment, 203139.38, leaving 434.17 as the sum of the squares due to varieties. Since there are 5 varieties there will be 4 degrees of freedom. The value 434.17, with the accompanying degrees of freedom, is recorded in line 1 of Table 95.

The variation between fields is obtained similarly by squaring the field totals and dividing the sum of the squares by 100, since each field total is derived from 5 varieties and 5 blocks during 4 years. We have 20356570/100, or 203565.70, and from this is subtracted the general correction, 203139.38, leaving 426.32 as the sum of the squares for fields. There is 1 degree of freedom, and the value, with the accompanying degree of freedom, is recorded in line 2 of Table 95.

For the interaction of varieties and fields, the individual total yields are squared and summed, giving 4090404, and this sum is divided by 20, since these individual total yields are derived from 5 blocks during 4 years. The general correction, 203139.38, is deducted from the quotient, 204520.20, leaving 1380.82. There are 9 degrees of freedom. It is necessary to take from this remainder, and the degrees of freedom, the sums of the squares and the degrees of freedom for the varieties and fields.

The steps just outlined for this analysis are given in detail on page 385. In order that it may be clear why it is necessary to subtract the sums of the squares for varieties and fields, it may be well to consider the analysis of interaction with that of a simple experiment. Referring to the simple analysis for the test in 1928, it is recalled that the effect of blocks and varieties was eliminated from the sum of the squares for total to obtain the error. The sum of the squares for total had been obtained by squaring the individual yields. For the analysis of the interaction, say of varieties and fields, a table is prepared giving all the yields of each variety for all the plots and years, so that the individual items are these totals rather than individual plot yields. In the study of the interaction the same process of analysis is followed as for a simple experiment, namely, the sum of the squares for total is obtained and from this is subtracted the sums of the squares due to varieties and fields, leaving the sum of the squares due to error, which in this case is the interaction between varieties and fields.

BETWEEN VARIETIES	BETWEEN FIELDS	INTERACTION
1503076 1452025 1871424 1597696 1718721 SUM 8142942 $\frac{8142942}{40}=203573.55$ 203573.55 -203139.38 434.17	11108889 9247681 SUM 20356570 $\frac{20356570}{100}=203565.70$ 203565.70 -203139.38 426.32	391876 469225 462400 405769 497025 360000 270400 473344 393129 367236 SUM 4090404 $\frac{4090404}{20}=204520.20$ 204520.20 -203139.38 1380.82

Sum of Squares for Total –Sum of Squares for Varieties –Sum of Squares for Fields	1380.82 –434.17 –426.32	9 Degrees of Freedom –4 Degrees of Freedom –1 Degree of Freedom
Difference	520.33	4 Degrees of Freedom

The difference obtained, 520.33, is the sum of the squares for the interaction of varieties and fields, and with its accompanying degrees of freedom is recorded in line 5 of Table 95.

For the variation due to varieties and years and their interaction, the yields are combined as shown at the top of page 386. These individual values are obtained by combining the total of each variety on Field I with the total for that variety on Field II. For example, for variety A in 1928 the total yield on Field I is 104 and on Field II it is 122, and the sum is 226 as recorded. For variety

TOTAL YIELDS FOR VARIETIES AND YEARS

YEARS	VARIETIES					TOTAL
	A	B	C	D	E	
1928	226	281	304	199	216	1226
1929	243	255	263	228	256	1245
1930	338	283	397	410	464	1892
1931	419	386	404	427	375	2011
TOTAL	1226	1205	1368	1264	1311	6374

B in 1928 we have a total yield of 137 on Field I and of 144 on Field II, and the sum is 281. The remaining yields are similarly obtained.

The variation due to varieties has already been determined, as recorded in line 1 of Table 95. For the variation due to years, the totals are squared, summed, and divided by 50, since each total is derived from 5 blocks of 5 varieties on 2 fields. We have 10676886/50, or 213537.72. From this is deducted the general correction, 203139.38, leaving 10398.34, which is the sum of the squares for years. This value, with its accompanying degrees of freedom, is recorded in line 3 of Table 95.

For the interaction of varieties and years, the individual total yields are squared, summed, and divided by 10, since each individual total yield is made up of 5 blocks on 2 fields. We have 2165538/10, or 216553.80. From this is subtracted the general correction, 203139.38, leaving 13414.42 with 19 degrees of freedom. We must deduct from this remainder, and the degrees of freedom, the sums of squares and degrees of freedom for varieties and years. We have

Sum of Squares for Total	13414.42	19 Degrees of Freedom
–Sum of Squares for Varieties	– 434.17	–4 Degrees of Freedom
–Sum of Squares for Years	–10398.34	–3 Degrees of Freedom
Difference	2581.91	12 Degrees of Freedom

This value, 2581.91, is the sum of the squares for the interaction of varieties and years, and with its accompanying degrees of freedom is recorded in line 6 of Table 95.

For the variation due to fields and years and their interaction the yields are combined as follows:

TOTAL YIELDS FOR FIELDS AND YEARS

YEARS	FIELDS		TOTAL
	I	II	
1928	566	660	1226
1929	659	586	1245
1930	1049	843	1892
1931	1059	952	2011
TOTAL	3333	3041	6374

The individual yields are obtained by combining the yields of all varieties and blocks for one year on each field, or these values are the same as the totals for the individual tests in Table 94.

Since the variation due to fields and years has already been determined and recorded in lines 2 and 3 of Table 95, it is necessary to consider only their interaction. The individual total yields are squared, summed, and divided by 25, since each individual total yield is made up of 5 blocks of 5 varieties. We have 5372468/25, or 214898.72, and from this must be deducted the general correction, 203139.38. The remainder is 11759.34, and there are 7 degrees of freedom. It is necessary to deduct from this remainder the sums of squares and degrees of freedom for fields and years and we have

Sum of Squares for Total	11759.34	7 Degrees of Freedom
–Sum of Squares for Fields	– 426.32	–1 Degree of Freedom
–Sum of Squares for Years	–10398.34	–3 Degrees of Freedom
Difference	934.68	3 Degrees of Freedom

This value, 934.68, is the sum of the squares for the interaction of fields and years, and with its accompanying degrees of freedom is recorded in line 7 of Table 95.

The interactions that have been obtained up to this point are known as the first order interactions, that is, they are the interactions between two variables. We may now consider the interactions between three variables, or the second order interactions. If it is necessary to determine further interactions, similar methods are followed for other relations.

For the second order interaction of varieties, fields, and years, the yields are combined as follows:

TOTAL YIELDS FOR VARIETIES, FIELDS, AND YEARS

FIELDS	VARIETIES					TOTAL
	A	B	C	D	E	
1928						
I	104	137	125	87	113	566
II	122	144	179	112	103	660
1929						
I	128	131	135	127	138	659
II	115	124	128	101	118	586
1930						
I	171	196	225	207	250	1049
II	167	87	172	203	214	843
1931						
I	223	221	195	216	204	1059
II	196	165	209	211	171	952
TOTAL	1226	1205	1368	1264	1311	6374

These individual values are the variety totals recorded in Table 94.

Since the variation due to varieties, fields, and years, and their interaction, has been determined and recorded in Table 95, it is necessary to consider only the second order interaction. The individual total yields are squared, summed, and divided by 5, since each individual total yield is made up of 5 blocks. We have 1096044/5, or 219208.80. From this must be deducted the general correction, 203139.38, leaving 16069.42 with 39 degrees of freedom. From this must be deducted the sum of the squares and degrees of freedom for varieties, fields, and years, and all first order interactions of these three factors. In other words, the sums of the squares and degrees of freedom determined thus far in the complex experiment must be deducted. We have

Sum of Squares for Total	16039.42	39 Degrees of Freedom
–Sum of Squares for Varieties	– 434.17	– 4 Degrees of Freedom
–Sum of Squares for Fields	– 426.32	– 1 Degree of Freedom
–Sum of Squares for Years	–10398.34	– 3 Degrees of Freedom
–Sum of Squares for Interaction Varieties × Fields	– 520.33	– 4 Degrees of Freedom
–Sum of Squares for Interaction Varieties × Years	– 2581.91	–12 Degrees of Freedom
–Sum of Squares for Interaction Fields × Years	– 934.68	– 3 Degrees of Freedom
Difference	773.67	12 Degrees of Freedom

This value, 773.67, is the sum of the squares for the second order interaction of varieties, fields, and years, and with its accompanying degrees of freedom is recorded in line 8 of Table 95.

For the variation due to blocks, and the interaction of blocks and the other factors, the same plan is followed. For blocks and years the yields are combined as shown on page 390. These individual total yields are obtained by combining the total block yields on the two fields for each year. For example, for block 1 in 1928 we have 129 + 125, or 254, and the other individual total yields are similarly obtained.

The variation due to years has already been determined and is recorded in line 3 of Table 95. For the variation due to blocks the block totals are squared, summed, and divided by 40, since each block total is made up of 5 varieties on 2 fields for 4 years. We have 8137030/40, or 203425.75, and from this must be deducted the general correction, 203139.38. The remainder, 286.37 is the sum of the squares for blocks, and with its accompanying degrees of freedom is recorded in line 4 of Table 95.

TOTAL YIELDS FOR BLOCKS AND YEARS

YEARS	BLOCKS					TOTAL
	1	2	3	4	5	
1928	254	232	220	278	242	1226
1929	244	236	273	268	224	1245
1930	323	392	457	349	371	1892
1931	466	430	364	407	344	2011
TOTAL	1287	1290	1314	1302	1181	6374

For the interaction due to blocks and years, the individual total yields are squared and summed and divided by 10, since each individual total yield is made up of 5 varieties on 2 fields. We have 2159170/10, giving 215917.00, and from this is deducted the general correction, 203139.38. The remainder is 12777.62, with 19 degrees of freedom. From this must be deducted the sums of the squares and degrees of freedom for blocks and years, and we have

Sum of Squares for Total	12777.62	19 Degrees of Freedom
–Sum of Squares for Blocks	– 286.37	– 4 Degrees of Freedom
–Sum of Squares for Years	–10398.34	– 3 Degrees of Freedom
Difference	2092.91	12 Degrees of Freedom

This value, 2092.91, is the sum of the squares for the interaction of blocks and years, and with its accompanying degrees of freedom is recorded in line 9 of Table 95.

For the interaction of blocks and fields the yields are combined as follows:

TOTAL YIELDS FOR BLOCKS AND FIELDS

FIELDS	BLOCKS					TOTAL
	1	2	3	4	5	
I	596	683	711	720	623	3333
II	691	607	603	582	558	3041
TOTAL	1287	1290	1314	1302	1181	6374

These individual total yields are obtained by summing the total of each block for the four years of the test. For example, for block 1 on Field I we have 129+115+121+231, or 596, and the other individual total yields are similarly obtained.

The variation due to blocks and fields has already been determined and recorded in lines 4 and 2 of Table 95. For the interaction of these two factors, each individual total yield is squared, summed, and divided by 20, since each individual total yield is made up of 5 varieties for 4 years. We have 4093382/20, giving 204669.10, and from this must be deducted the general correction, 203139.38. The remainder is 1529.72 with 9 degrees of freedom, and from this must be deducted the sums of the squares and degrees of freedom for blocks and fields. We have

Sum of Squares for Total	1529.72	9 Degrees of Freedom
–Sum of Squares for Blocks	– 286.37	–4 Degrees of Freedom
–Sum of Squares for Fields	– 426.32	–1 Degree of Freedom
Difference	817.03	4 Degrees of Freedom

This value, 817.03, is the sum of the squares for the interaction of blocks and fields, and with its accompanying degrees of freedom is recorded in line 10 of Table 95.

For the second order interaction of blocks, fields, and years, the yields are combined as follows:

TOTAL YIELDS FOR BLOCKS, FIELDS, AND YEARS

FIELD I	BLOCKS					TOTAL
	1	2	3	4	5	
1928	129	108	92	127	110	566
1929	115	122	151	154	117	659
1930	121	214	272	225	217	1049
1931	231	239	196	214	179	1059
FIELD II						
1928	125	124	128	151	132	660
1929	129	114	122	114	107	586
1930	202	178	185	124	154	843
1931	235	191	168	193	165	952
TOTAL	1287	1290	1314	1302	1181	6374

These individual values are the block totals recorded in Table 94.

Since the variation due to blocks, fields, and years, and their interaction, has been determined and recorded in Table 95, it is necessary to consider only their second order interactions. The individual total yields are squared, summed, and divided by 5, since each individual total yield is made up of 5 varieties. We have 1098968/5, giving 219793.60, and from this must be deducted the general correction, 203139.38. The remainder is 16654.22, with 39 degrees of freedom, and from this remainder must be deducted the sums of squares and degrees of freedom for blocks, fields, and

Sum of Squares for Total	16654.22	39 Degrees of Freedom
–Sum of Squares for Blocks	– 286.37	– 4 Degrees of Freedom
–Sum of Squares for Fields	– 426.32	– 1 Degree of Freedom
–Sum of Squares for Years	–10398.34	– 3 Degrees of Freedom
–Sum of Squares for Interaction Fields × Years	– 934.68	– 3 Degrees of Freedom
–Sum of Squares for Interaction Blocks × Years	– 2092.91	–12 Degrees of Freedom
–Sum of Squares for Interaction Blocks × Fields	– 817.03	– 4 Degrees of Freedom
Difference	1698.57	12 Degrees of Freedom

TABLE 95

ANALYSIS OF VARIANCE FOR ENTIRE EXPERIMENT

	VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE	*F* VALUE FOR COMPARISON
1	Varieties	4	434.17	108.5425	3.22
2	Fields	1	426.32	426.3200	12.66
3	Years	3	10398.34	3466.1133	102.93
4	Blocks	4	286.37	71.5925	2.13
	Interaction				
5	Varieties × fields	4	520.33	130.0825	3.86
6	Varieties × years	12	2581.91	215.1592	6.39
7	Fields × years	3	934.68	311.5600	9.25
8	Varieties × fields × years	12	773.67	64.4725	1.91
9	Blocks × years	12	2092.91	174.4092	5.18
10	Blocks × fields	4	817.03	204.2575	6.07
11	Blocks × fields × years	12	1698.57	141.5475	4.20
12 Error	Blocks × varieties	16 TOTAL 128	4310.32	33.6744	
	Blocks × varieties × fields	16			
	Blocks × varieties × years	48			
	Blocks × varieties × fields × years	48			
13	TOTAL	199	25274.62		

years, and all first order interactions of these factors. These sums of squares and degrees of freedom are obtained from Table 95 and are recorded at the bottom of page 392. This value, 1698.57, is the sum of the squares for the interaction of blocks, fields, and years, and with its accompanying degrees of freedom is recorded in line 11 of Table 95, page 393.

To obtain the variation due to error, the sums of the squares in lines 1 to 11 of Table 95 are added, giving 20964.30 This is subtracted from the total sum of the squares, 25274.62, in line 13, leaving 4310.32 as the variation due to error. It is noted that this difference includes the interaction between blocks and varieties and all second order interactions where blocks and varieties are concerned. Dividing the error into its various components indicates the sums of squares that enter into the error, as well as the degrees of freedom for each component.

The amount left for error, 4310.32, may be checked by making a separate analysis of each field for each year. The results for Field I for 1928 were given earlier, and the sum of the squares for error was found to be 129.04. Making similar separate analyses for the other years the sums of the squares as given in Table 96, page 395, are obtained. The sum of the squares for the same comparisons from the complex experiment are also given.

If for each analysis the sums of the squares for blocks and for varieties are added together and subtracted from the sum of the squares for total, we would obtain the amount for error for each analysis. Since we are interested only in the analysis of the whole experiment we may add the sums of the squares for total, blocks, and varieties for each separate analysis, and subtracting the total of the sums for blocks and varieties from the total sum of the squares we obtain 4310.32 for error. This is the same value as was obtained from the analysis of the complex experiment by separating the experiment into its several components.

This furnishes a useful method for checking the accuracy of the work as it enables us to compare the amount left for error and to compare the degrees of freedom for error. Since the degrees of freedom for error for each of the separate experiments are 16 and there are 8 separate experiments, we have 16 × 8, or 128 degrees of freedom for error, which checks with the degrees of freedom as obtained from the complete analysis of the experiment.

Table 96

Comparison of Sums of Squares Calculated from Individual Experiments and from the Analysis of the Complex Experiment

Sums of Squares Calculated Separately for Each Year

Field	Year	Total	Varieties	Blocks	Error
I	1928	609.76	295.36	185.36	
	1929	513.76	17.36	291.76	
	1930	4196.96	710.16	2410.96	
	1931	787.76	114.16	487.76	
II	1928	1190.00	738.80	98.00	
	1929	460.16	86.16	57.36	
	1930	3959.04	1983.44	735.04	
	1931	1797.84	364.64	628.64	
Sum		13515.28	4310.08	4894.88	4310.32

Sums of Squares

Variation Due to	From Individual Experiments	From Complex Experiment as Recorded in Table 95, in Lines
Total	13515.28	13515.28 (1,4,5,6,8,9,10,11,12)
Varieties	4310.08	4310.08 (1,5,6,8)
Blocks	4894.88	4894.88 (4,9,10,11)
Error	4310.32	4310.32 (12)

The results of the experiment may now be studied by comparing the variance for the separate components with the variance for error. The mean square, or variance, for each component is obtained by dividing the sum of the squares by the degrees of freedom. From these several mean squares the F values are determined by dividing the mean squares for the components by the mean square for error. Referring to Table X in the Appendix, it is seen that all of the comparisons show significance except the result from the blocks. The comparison of particular interest is that for varieties, and the value of F obtained lies between the two values given in Table X for the degrees of freedom concerned. It is nearer the higher value, so we may conclude with certainty that there is a difference in the yielding ability of the several varieties.

Since the varieties show significance when compared with the error of the experiment, we may now compare the varieties themselves. Extracting the square root of the variance for error, 33.6744, we have 5.80 for the error of a single plot. Dividing the variety totals by 40, the number of plots of each variety, we obtain the average variety yields as follows:

VARIETY	AVERAGE YIELD
A	30.65
B	30.12
C	34.20
D	31.60
E	32.77

For the standard error of 40 plots we divide the standard error of a single plot, 5.80, by $\sqrt{40}$, giving .92 as the standard error of the mean of 40 plots. It is necessary to obtain the standard error of the difference of two means, by multiplying the standard error of the mean of 40 plots by $\sqrt{2}$. We have

$$.92\sqrt{2}=1.30$$

Taking twice this value as being the smallest difference to be considered significant, we have 2.60.

Using 2.60 as the standard, it is seen that variety C is better than varieties A and B. The difference between variety C and variety D is 2.60, so it is hardly safe to conclude that variety C is better than variety D. Variety E may be considered slightly better than variety B.

It is apparent that by eliminating the effect of fields and years the error has been greatly reduced. The variances for the interactions are significantly higher than the variance for error, as may be determined from the values of F in Table X. It, therefore, is apparent that there is some difference in response by the varieties to the different years or fields.

Since the method of the analysis of a complex experiment is an extension of the method of variance analysis, the steps have been given in considerable detail. This method shows the advantage of separating a large experiment into its several components, and it is possible to eliminate some less important information and determine the factors of more importance. It is useful in an experiment of this sort where several varieties are compared in different localities and for different years, and it is also convenient for the analysis of experiments from soil fertility studies, rotation experiments, and the like. If through accident or otherwise the results from a certain plot or plots have been lost, it is possible to calculate the yields of the missing plots by the methods already explained and then complete the analysis.

CHAPTER XV

PROBLEMS OF PLOT TECHNIC

Since the methods of statistical analysis as presented in this volume are particularly applicable to problems in agricultural research and since there will be considerable interest on the part of the readers for information on plot technic, it is thought best to present as the last chapter a short discussion dealing with this particular phase of the subject. It is not possible in this chapter to discuss in detail all of the problems involved in plot layout and the interpretation of results, but the important points will be discussed briefly.

Plot technic has to do with the study of the kinds of plots best adapted for a particular type of experimental analysis. That is, one uses field plots for the purpose of determining the yielding ability of different kinds or varieties of plants, the effect of different cultural methods on plant growth, or the effect of different kinds and amounts of fertilizer. There may be other purposes for which one uses field plots in the study of certain problems, but in general the studies have to do with variety testing, cultural and rotation experiments, and the effect of fertilizers. It is important for each particular experiment that the best type and arrangement of plots be used. The purpose of this chapter is to discuss the kinds of plots best suited for a special type of work, the ways of determining the most suitable plots, and methods for the interpretation of the results obtained.

In the earlier history of plot experiments it was thought sufficient to have one plot of each variety or kind of treatment, but the results of these studies were rather variable and often conflicting. This led to a more careful study of field plots and it was found that although a field was level and apparently uniform, yet from the standpoint of fertility the different parts of the field varied greatly. Even adjacent plots did not produce exactly the same results. It was therefore recognized that since plots varied so greatly it would be necessary to use more than one plot to measure the yielding capacity of a variety or the effect of a fertilizer. As the variation in plots on the same field came to be recognized, studies were made to determine the extent of such variation.

Soil Heterogeneity. Harris was one of the first to study plot variation on a large scale, and he developed the method for determining the coefficient of heterogeneity, or the coefficient which determines the extent of plot variability. This method consists of making special application of the principles of correlation, and may be illustrated by the data in Table 97, page 400.

For this analysis we have assumed yields for 32 plots, and by Harris' method the plots may be grouped in various ways. For example, we may have a grouping of 1 by 2, which means one plot in width and two plots in length, and a 2 by 1 combination would be two plots in width and one in length. We may arrange any other combination, such as 1 by 4 or 4 by 1, and so on. A 2 by 2 combination is used for the analysis in the present case, or a grouping of two plots in width and two plots in length. The necessary values to be determined are as follows:

M_p = Average of all individual plots
n = Number of plots in each group
m = Number of groups of plots
$\Sigma\,(P^2)$ = Sum of the squares of the yields of the individual plots
$\Sigma\,(C_p^2)$ = Sum of the squares of the total yield of each group
σ_p^2 = Variance of the yields of the individual plots

The formula for the coefficient of heterogeneity is

$$\text{Coefficient of heterogeneity, or } r = \frac{\{[\Sigma(C_p^2) - \Sigma(P^2)]/m[n(n-1)]\} - M_p^2}{\sigma_p^2}$$

It will be noted that the mean yield of all plots is needed, as well as the sum of the squares of the yields of the individual plots and the sum of the squares of the total yield of the combination plots. For the combination plot for the first grouping we have the sum of the yields of the four plots, which are 22, 23, 23, and 22. Their sum is 90, and the square of this sum is 8100. The squares of the other combination plots are obtained in the same way, and the

TABLE 97

APPLICATION OF THE METHOD FOR DETERMINING THE COEFFICIENT OF HETEROGENEITY

22	23	22	20
23	22	22	20
21	22	21	19
22	21	21	19
20	18	20	19
18	20	19	20
18	17	20	18
18	17	20	18

M_p = Average of all individual plots = 20

n = Number of plots in each group = 4

m = Number of groups of plots = 8

$\Sigma(P^2)$ = Sum of the squares of the yields of the individual plots = 12892

$\Sigma(C_p^2)$ = Sum of the squares of the total yield of each group = 51488

σ_p^2 = Variance of the yields of the individual plots = 2.8750

C_p	C_p^2	P	P^2	DEVIATION FROM MEAN	D^2
90	8100	22	484	2	4
86	7396	23	529	3	9
76	5776	23	529	3	9
70	4900	22	484	2	4
84	7056	21	441	1	1
80	6400	22	484	2	4
78	6084	22	484	2	4
76	5776	21	441	1	1
		20	400	0	0
$\Sigma(C_p^2) = 51488$		18	324	−2	4
		18	324	−2	4
		20	400	0	0

etc. etc.

$\Sigma(P^2) = 12892$

$\Sigma D^2 = 92$

$N = 32$

$\sigma_p^2 = 92/32 = 2.8750$

Coefficient of heterogeneity, or
$$r = \frac{\{[\Sigma(C_p^2) - \Sigma(P^2)]/m[n(n-1)]\} - M_p^2}{\sigma_p^2}$$

$$r = \frac{\{[(51488 - 12892)]/8\ [4(4-1)]\} - 400}{2.8750} = \frac{2.0417}{2.8750} = .710$$

total sum obtained. The variance is determined by the usual method of obtaining the deviation of each individual plot yield from the mean of all the plots. The various steps are followed in Table 97 and the values substituted in the formula. The coefficient of heterogeneity, or *r*, for which the probable error is calculated as usual, is .710±.059, which indicates a high relation between the different plots and a high heterogeneity factor. If the yields of adjacent plots were not related the coefficient of heterogeneity would be low.

Harris made many studies to determine whether soil variation is a general condition or limited to certain fields or environments. His studies covered a wide range of crops and fields from many different localities, which meant that the studies were made under varying environmental conditions. A few of the results obtained by Harris are presented in Table 98.

TABLE 98

PRACTICAL UNIVERSALITY OF FIELD HETEROGENEITY AS SHOWN BY THE COEFFICIENTS OF CORRELATION AND PROBABLE ERRORS FOR VARIOUS COMBINATIONS OF PLOTS

CROP	COEFFICIENT OF CORRELATION
Potatoes	.311 ± .043
Timothy hay	.611 ± .027
Grain in wheat	.336 ± .027
Straw in wheat	.483 ± .023
Kherson oats	.495 ± .035
Grain content in wheat	.391 ± .038
Hops	
1909	.444 ± .099
1910	.695 ± .064
1911	.061 ± .123
1912	.326 ± .110
1913	.606 ± .078
1914	.386 ± .105
Unhusked rice	.344 ± .081
Ear corn	
1895	.830 ± .019
1896	.815 ± .021
1897	.606 ± .039

These results indicate that there is a tendency for fields to show a rather high degree of soil variation or heterogeneity. Occasionally fields are found which show a fairly low coefficient but the studies by Harris and others indicate a tendency for most fields to vary considerably. Since the studies have indicated that there is considerable variability between plots, it is important to know whether this variability is permanent, or, in other words, whether there is any tendency for plots that produce low yields in one season to produce low yields the following season, and whether plots that yield high in one season will also yield high the next season. Harris and Scofield made extensive studies in this connection and found that while there is some variation so that the plots in all cases do not tend to remain high or low, in general this tendency holds. Some of the results obtained by Harris and Scofield are presented in Tables 99 and 100.

TABLE 99

INTERANNUAL CORRELATIONS FOR YIELD OF HOPS

BEGINNING OF SERIES	FIRST AND SECOND YEARS	FIRST AND THIRD YEARS	FIRST AND FOURTH YEARS	FIRST AND FIFTH YEARS	FIRST AND SIXTH YEARS
1909	.768 ± .051	.622 ± .075	.380 ± .105	.259 ± .115	.061 ± .123
1910	.577 ± .082	.447 ± .099	.451 ± .098	.274 ± .114	
1911	.062 ± .123	.313 ± .111	−.126 ± .1 1		
1912	.311 ± .111	.705 ± .062			
1913	.597 ± .079				

In Table 99 three out of five of the correlations between the first and second years are significant; three out of four between the first and third years are significant; and two out of three for the first and fourth years are significant. There is less correlation between the yields for the first and fifth and the first and sixth years. The results in this table indicate that there is a tendency for the plots to yield in a similar manner from year to year, although there are some exceptions. The results in Table 100, page 403, showing the yields obtained from alfalfa on the same plots for different cuttings or harvests, all show significant correlations.

TABLE 100

COMPARISON OF CORRELATIONS BETWEEN DIFFERENT CUTTINGS OF ALFALFA IN THE SAME YEAR

CUTTINGS OF ALFALFA	WHOLE PLOTS	HALF PLOTS	QUARTER PLOTS
1913, First and Second Cuttings	.454 ± .079	.442 ± .057	
1914, First and Second Cuttings	.711 ± .049	.633 ± .042	.558 ± .034
1914, First and Third Cuttings	.595 ± .064		
1914, (First plus Second) and Third Cuttings	.653 ± .057		

Similar results have been found by other investigators. Parker and Batchelor, working with citrus trees, obtained results as shown in Table 101.

TABLE 101

INTERANNUAL CORRELATION COEFFICIENTS FOR YIELDS OF INDIVIDUAL TREES

	1922	1923	1924	1925	1926	1927
1921	.637 ± .010	.260 ± .016	-.173 ± .017	.170 ± .017	-.171 ± .017	-.083 ± .017
1922		.307 ± .016	.324 ± .016	.455 ± .014	.069 ± .017	.153 ± .017
1923			.595 ± .011	.415 ± .014	.347 ± .015	.255 ± .016
1924				.685 ± .009	.532 ± .012	.468 ± .014
1925					.550 ± .017	.488 ± .013
1926						.536 ± .012

It is seen that with a few exceptions there is a general tendency for the plots to yield similarly from year to year. Similar results have been obtained by Garber, McIlvaine, and Hoover, as shown in Table 102.

TABLE 102

CORRELATIONS BETWEEN THE SAME PLOTS IN DIFFERENT YEARS AND FOR DIFFERENT CROPS

CORRELATION BETWEEN	N	r
Corn (grain) 1927 and oats (grain) 1928	445	.58 ± .02
Corn (stover) 1927 and oats (straw) 1928	411	.70 ± .02
Corn (grain) 1927 and wheat (grain) 1929	445	.20 ± .03
Corn (stover) 1927 and wheat (straw) 1929	412	.33 ± .03
Oats (grain) 1928 and wheat (grain) 1929	445	.50 ± .02
Oats (straw) 1928 and wheat (straw) 1929	412	.49 ± .03

With the fact of plot variability in mind it is evident that it is not wise to use the data from single plots for measuring results, but that it is necessary to have several plots treated in a similar manner. Where the yield of a certain variety of grain or the effect of a certain treatment is desired, several plots must be sown to the same variety or treated similarly. This leads to the question of the number of plots to be used for any particular study, and necessitates consideration of the size and shape of plots.

Size, Shape, and Replication of Plots. There are a number of factors to be considered in determining the size, shape, and replication of plots. Some of these are mechanical factors, such as the matter of preparing the plots and their planting and cultivation. If all of the preparation and handling of the plots is to be done by hand, then the plots may be of a different size and shape than if machinery is to be used in connection with the work of preparation, planting, and harvesting. The amount of land available for the experiment is also a determining factor.

Another factor to be considered in connection with the size of plots is the kind of crop that is to be grown. If the crop is of such nature that each plant requires a considerable amount of space, then the plots must be larger in order to have a number of individual plants on the same plot of ground. With extremely large plants, such as trees, it is never possible to have a large number on any one plot. With other crops, like cotton, corn, or kaoliang, it is possible to have plots large enough to include a number of plants on each plot. Recommendations as to the size of plots for such crops will be made after considering certain other problems.

Various studies have been made to determine the best size and shape of plot. These studies are based on the effect of different sizes and shapes of plots on either the probable or standard error, and in this connection the errors for the various sizes and shapes have been determined. The results of some of the earlier investigations, presented by Hall and Russell from studies of various plot arrangements, are given in Table 103.

TABLE 103

PROBABLE ERRORS IN PER CENT OBTAINED FROM PLOTS DIFFERING IN SIZE, AND FROM PLOTS MADE UP OF SEVERAL SCATTERED UNITS

SIZE OF PLOT (*acres*)	PROBABLE ERROR (*per cent*)	
1/500	7.8	Simple plots *
1/250	6.7	Simple plots
1/125	6.0	Simple plots
1/50	4.2	Simple plots
1/25	3.8	Simple plots
1/10	3.4	Simple plots
1/100	3.1	Plots made up of 5 scattered units
1/50	2.4	Plots made up of 5 scattered units
1/10	1.6	Plots made up of 5 scattered units
1/5	1.3	Plots made up of 5 scattered units
1/5	3.1	Single plot
1/5	1.7	Made up of 2 scattered 1/10
1/5	1.3	Made up of 5 scattered 1/25
1/5	1.1	Made up of 10 scattered 1/50

* This has been interpreted as meaning single plots.

These results are obtained from dividing a field into plots of 1/500 acre each. By combining the smaller plots it is then possible to make up plots of various sizes. Results are given for single plots and also for various sized plots made up by combining scattered units. These data show that the probable error in per cent decreases as the size of the plot increases. The probable error of a single observation for a plot 1/500 of an acre is 7.8 per cent; for a plot 1/125 of an acre the probable error is 6.0 per cent; and for a plot 1/10 of an acre the probable error is only 3.4 per cent. A single plot 1/50 of an acre has a probable error of 4.2 per cent, but if a plot of the same size is made up by combining scattered units the probable error of a single observation is reduced to 2.4 per cent. That is, with the same area of land but using smaller plots replicated four times the probable error of a single observation is greatly reduced. A single plot 1/5 of an acre has a probable error of 3.1 per cent, but when made up of scattered units 1/10, 1/25, or 1/50 of an acre the probable error is greatly reduced.

Data obtained by McClelland from field plot trials with corn are of interest in this connection, and are given in Table 104.

TABLE 104

PROBABLE ERRORS FOR SINGLE PLOTS IN PER CENT, OBTAINED FROM PLOTS VARYING IN SIZE AND SHAPE

SIZE OF PLOTS	NUMBER OF PLOTS	NUMBER OF ROWS 44″ APART	LENGTH OF ROWS IN FEET	ERROR %
1/180 acre	432	1	66	11.2
1/90 acre	216	2	66	10.3
1/90 acre	216	1	132	9.7
1/60 acre	144	3	66	9.7
1/45 acre	108	4	66	9.7
1/45 acre	108	2	132	9.0
1/36 acre	86	5	66	9.5
1/30 acre	72	6	66	9.4
1/30 acre	72	3	132	8.6
1/20 acre	48	9	66	8.9
1/18 acre	42	10	66	9.0
1/18 acre	43	5	132	9.6
1/15 acre	36	12	66	8.8
1/15 acre	36	6	132	8.6
1/12 acre	28	15	66	8.6
1/10 acre	24	18	66	8.4
1/10 acre	24	9	132	8.3
1/5 acre	12	36	66	6.8
1/5 acre	12	18	132	7.4
1/2 acre	4	90	66	6.95
1/2 acre	4	45	132	6.2

Plots of various sizes from 1/180 to 1/2 of an acre were used. In general there is a decrease in the probable error of a single observation in per cent as the size of the plot increases. When plots of the same size are compared the probable error of a single observation is slightly smaller in general for the long narrow plots than for the short wide plots, although there are exceptions in some cases.

Further data on the size and shape of plots are given by Day, from results obtained from a field of wheat. These results are given in Table 105, page 408. These data show first that as the length of row, or plot, increases there is a general tendency for the variability to decrease. Relative to the width of plot, when three adjacent 50-foot rows are used as one plot the coefficient of variability is 16.37, while for five adjacent rows it is reduced to 14.49. As the width of plot increases up to 20 adjacent rows there is a decrease in the variability. The same reduction in variability is true for rows 15 feet in length when 10 or 20 adjacent rows are compared with 5 adjacent rows.

The data in the lower part of Table 105 show also that when the plots have their greater length in the direction of the greatest variation the variability is reduced. For example, in the comparison of plots of equal area (total length of row) for 15 adjacent rows 50 feet in length and 50 adjacent rows 15 feet in length the coefficient of variability is 10.18 when the plot is long in the direction of least variation, while the coefficient of variability is only 7.45 when the plot is long in the direction of the most variation. Other comparisons may be made.

The effect of replication is also shown by the data in Table 106, page 409, which have been presented by Day. In general these data show that as the number of replications increase there is a reduction in variability.

TABLE 105

COEFFICIENTS OF VARIABILITY OBTAINED FROM PLOTS OF DIFFERENT SIZES AND SHAPES

SIZE OF PLOT		COEFFICIENT OF VARIABILITY
1	5-foot row	37.20
1	10-foot row	29.58
1	15-foot row	26.52
1	20-foot row	24.41
1	25-foot row	22.81
1	30-foot row	22.53
1	35-foot row	21.32
1	40-foot row	20.28
1	45-foot row	19.85
1	50-foot row	20.67
1	60-foot row	18.99
1	75-foot row	19.64
1	100-foot row	16.74
1	125-foot row	17.01
1	150-foot row	17.36
3	adjacent 50-foot rows	16.37
5	adjacent 50-foot rows	14.49
10	adjacent 50-foot rows	12.13
15	adjacent 50-foot rows	10.18
20	adjacent 50-foot rows	8.32
5	adjacent 15-foot rows	16.49
10	adjacent 15-foot rows	12.72
20	adjacent 15-foot rows	9.98

UNITS OF WHICH PLOT IS COMPOSED

NO. OF ADJACENT ROWS	LENGTH OF ROWS (*feet*)	LENGTH OF ROWS IN PLOT (*feet*)	SHAPE OF PLOT	COEFFICIENT OF VARIABILITY
3	50	150	Long in direction of least variation	16.37
10	15	150	Rectangular	12.72
24	5	120	Long in direction of most variation	10.54
5	155	775	Long in direction of least variation	13.07
15	50	750	Long in direction of least variation	10.18
50	15	750	Long in direction of most variation	7.45
10	155	1550	Long in direction of least variation	9.43
30	50	1500	Somewhat long in direction of least variation	5.46
100	15	1500	Long in direction of most variation	2.77

TABLE 106

COEFFICIENTS OF VARIABILITY OBTAINED FROM REPLICATING PLOTS OF DIFFERENT SIZES

COMPOSITION OF BLOCK UNIT			
NO. OF BLOCKS	NO. OF ADJACENT ROWS	LENGTH OF ROWS (*feet*)	COEFFICIENT OF VARIABILITY
5	3	50	3.97
10	3	50	3.35
14	3	50	3.50
5	5	50	2.94
10	5	50	2.38
7	5	15	5.04
14	5	15	3.32
28	5	15	1.83
3	10	15	5.97
7	10	15	3.75
14	10	15	1.63
3	20	15	4.51
6	20	15	1.27
3	50	15	5.29
5	8	5	7.53
10	8	5	4.35

The foregoing results indicate that as plots are increased in size there is a general tendency for the variability to decrease. The variability is less when a unit of a certain area is made up of several distributed units than when a single larger unit is used. This question will be considered further in the discussion of blank tests.

Blank Tests. In recent years considerable attention has been given to the study of the proper size, shape, and replication of plots, and this has led to the analysis of many blank tests. By a blank test is meant that a large field is given uniform treatment and sown to the same variety of crop in such a way that plots of different sizes and shapes may be made up by a combination of various groups of small plots for the purpose of determining the best type of plot and the proper number of replications for a particular experiment. For example, if one were to study the size and shape of plots for a crop such as wheat, the field would be sown to a large number of individual rows, having the rows one foot apart and long enough so that rows of different lengths could be obtained by harvesting the actual row in small units. For example, the first unit may be eight feet in length and the remainder of each row may be harvested in units four feet in length, or if one chooses the units may be two feet in length, thus making it possible to make up rows of various lengths. With such a planting arrangement it is also possible to have plots of various widths by combining two or more rows. With such crops as cotton a similar planting arrangement may be followed, allowing the proper distance between the rows. It is possible to arrange blank tests for the different kinds of crops, so that one may determine for a particular crop or environment the best size or shape of plot and the number of replications to use.

The analysis of the data may be based on the probable error or on the variance, depending on the method that one chooses to follow. One method that has been frequently used is to determine the probable error for plots of different sizes, and the plot that gives the smallest error may be considered the most reliable. Whether it is the one best suited for the actual experiments will depend on other considerations, such as ease in handling and the amount of land required, as well as on the number of different variables that are to be determined.

As an illustration of the results from a blank test, we may consider the analysis of the data from a blank test of cotton. For this study the seeds were planted in rows 100 feet long, and at harvest time the first 20 feet of each row were harvested separately. The remainder of each row was then harvested in 5-foot units, making it possible to have rows 20, 25, 30, 35, and 40 feet in length. By adding a 5-foot unit it would be possible to continue to any length by 5-foot units up to 100 feet. Part of the data giving the weights for different lengths of row obtained from this field are shown in Table 107.

TABLE 107

YIELDS OBTAINED IN A BLANK TEST OF COTTON

	LENGTH OF UNITS HARVESTED																
Row No.	20 ft.	5 ft.	5 ft.	5 ft.	5 ft.	5 ft.	5 ft.	5 ft.	5 ft.	5 ft.	5 ft.	5 ft.	5 ft.	5 ft.	5 ft.	5 ft.	5 ft.
1	14.8	2.7	2.4	3.8	2.9	3.4	2.8	4.5	2.6	3.2	1.7	1.7	2.7	3.0	2.7	3.8	3.4
2	15.2	4.3	4.7	2.3	1.5	3.8	4.0	4.0	3.4	3.4	2.1	3.2	2.7	3.0	4.9	2.3	1.4
3	13.4	2.7	3.3	3.2	3.7	3.2	2.2	4.4	3.0	3.6	2.8	1.4	2.8	2.0	2.6	3.4	3.1
4	12.7	2.9	3.5	2.2	3.9	3.6	3.0	3.4	2.9	2.4	2.1	1.7	2.6	3.0	3.2	2.9	3.6
5	13.0	3.4	3.6	1.9	3.2	3.3	2.9	4.3	2.8	2.1	1.7	2.8	2.8	3.4	3.4	2.4	2.1
6	13.0	2.2	2.7	2.2	2.2	2.5	1.4	3.0	1.3	1.9	2.0	1.6	1.9	2.1	2.1	2.4	1.7
7	8.5	2.7	2.8	2.5	2.6	2.0	1.7	1.9	2.6	2.6	2.6	2.1	3.4	2.4	3.0	2.0	1.4
8	12.3	3.0	3.0	1.4	2.9	2.7	2.6	3.6	1.9	1.8	3.7	4.1	2.1	2.8	2.8	2.9	1.1
9	12.5	2.9	3.1	3.6	2.5	1.7	2.8	2.4	1.7	2.3	3.6	2.7	3.0	3.7	3.6	1.3	2.0
10	14.9	3.7	3.3	2.8	1.9	3.0	3.3	2.2	2.0	2.4	3.8	3.0	2.5	3.4	4.0	1.9	1.7

Since these rows were all sown to the same variety of cotton it was possible to determine the probable error for each of the different lengths and widths of plots. This was done by determining the standard deviation for each type of plot, multiplying the standard deviation by the constant .6745 to give the probable error of a single determination, and expressing this as a percentage of the mean yield for the different lengths. The results obtained from the first analysis on the length of a single-row plot gave the following probable errors:

Length of plot	20 ft.	25 ft.	30 ft.	35 ft.	40 ft.
$P.E._s$ in per cent	12.56	11.97	11.34	10.68	10.09

These results show that as the length of plot increases there is a decrease in the probable error. The decrease for each added 5-foot unit is not large, but there is a tendency for the probable error to become gradually less until with the 40-foot unit it is 10.09 as compared with 12.56 for a 20-foot unit.

As indicated earlier, from a blank test it is possible to make up plots of different sizes and shapes, and the data from this blank test of cotton were used to make plots of various widths and lengths. The results of this analysis are given in Table 108, page 413.

Plots varying from a single row to seven rows in width, with the exception of the 6-row width, were studied. It is seen that in general the probable error decreases as the width of plot increases, and for the plots of various widths there is a general tendency for the probable error to decrease as the length of the plot increases. From such a blank test it is possible to study other sizes and shapes, but those presented here give some indication as to the reduction in variability as the size of the plot increases.

Blank tests may also be analyzed by the method for the analysis of variance. When using the method of variance we may assume that several varieties are distributed at random and then continue the replications until all the plots are used. In this study on the blank test of cotton the calculations are based on five varieties. The results will vary slightly, depending on the number of varieties assumed in the trial. With five varieties and the rows 20 feet long it is possible to have 139 replications, since there are 140 rows 100 feet long, giving 700 20-foot rows.

Table 108

Probable Errors for Single Plots in Per Cent, Obtained From Plots of Different Sizes

Probable Errors Obtained from Plots of Different Sizes

Width of Plot	Length of Plot in Feet				
	20	25	30	35	40
1 row	12.56	11.97	11.34	10.68	10.09
2 rows	10.35	9.74	9.61	9.05	8.40
3 rows	9.37	9.02	8.74	7.85	7.54
4 rows	9.51	8.93	8.67	7.96	7.29
5 rows	8.82	8.64	8.57	7.57	7.19
7 rows	8.36	7.86	7.75	7.04	6.73

Probable Errors Obtained by the Method of Variance Analysis

Width of Plot	Length of Plot in Feet		
	20	30	40
1 row	11.20	8.55	7.77
3 rows	7.87	5.10	6.56
5 rows	7.90	7.41	4.71

The steps in the analysis of variance are carried out in the usual manner. First, the total sum of the squares of the deviations of the individual plots from the mean is obtained. Then the total sum of the squares due to the difference between blocks is obtained. The difference between the total sum of the squares and the sum of the squares for the difference between blocks is determined, giving the sum of the squares due to the difference within blocks, or error. The degrees of freedom for error are obtained by subtracting the degrees of freedom for the variation between blocks from the degrees of freedom for total. By dividing the sum of the squares due to error by the appropriate degrees of freedom the variance for error is obtained. Extracting the square root and multiplying by the constant .6745 the probable error for a single plot is obtained.

The probable errors obtained for some of the combinations are given in Table 108. It will be observed that the probable errors obtained by the method of variance analysis are less than the probable errors obtained in the usual way. This is to be expected, since by the method of variance analysis we have eliminated the effect of variation between blocks, leaving only the variation due to the difference within blocks.

As already indicated, the size of plot to be used depends on other factors in addition to the actual size of the probable error. While in general the larger plots have a smaller probable error, it will be of interest to compare plots of different sizes on the basis of their efficiency in the use of land. The method is to take either the probable or the standard error of the smallest unit, for this particular illustration a single row 20 feet long, as a standard and assume that it is 100 per cent efficient in the use of land. Then plots of other sizes are compared with this on the basis of their probable error and the size of the plot as compared with the standard 20-foot plot. For example, the single-row 40-foot plot is twice the size of the 20-foot plot and the probable error for the 40-foot plot is 10.09. Squaring this probable error, 10.09, and multiplying by 2, the number of units of the 20-foot plot contained in the 40-foot plot, we have 203.6162. Dividing the square of the probable error of the standard 20-foot plot, $(12.56)^2$, by 203.6162 and multiplying by 100, we have 77.48 as the percentage of efficiency for the single-row 40-foot plot.

As another example, we may take a 30-foot plot four rows in width, which is equal to 120 feet of row (4×30) and contains six 20-foot units. The probable error for this plot is 8.67, and multiplying the square of this probable error by 6 gives 451.0134. Dividing the square of the probable error of the standard 20-foot plot, $(12.56)^2$, by 451.0134, we have 34.98 as the percentage of efficiency. On the basis of the 20-foot unit being 100 per cent efficient, the plot 30 feet long and four rows in width is 34.98 per cent as efficient. The efficiency of plots of various sizes and shapes has been calculated and the results are given in Table 109.

TABLE 109

EFFICIENCY IN THE USE OF LAND, EXPRESSED IN PER CENT, FOR PLOTS OF DIFFERENT SIZES, BASED ON THE DATA IN TABLE 108

BASED ON PROBABLE ERRORS OBTAINED FROM PLOTS OF DIFFERENT SIZES

WIDTH OF PLOT	LENGTH OF PLOT IN FEET				
	20	25	30	35	40
1 row	100.00	88.08	81.78	79.03	77.48
2 rows	73.63	66.52	56.94	55.03	55.89
3 rows	59.89	51.71	45.89	48.76	46.25
4 rows	43.61	39.56	34.98	35.57	37.11
5 rows	40.56	33.81	28.64	31.46	30.52
7 rows	32.25	29.18	25.01	25.98	24.88

BASED ON PROBABLE ERRORS OBTAINED BY THE METHOD OF VARIANCE ANALYSIS

WIDTH OF PLOT	LENGTH OF PLOT IN FEET		
	20	30	40
1 row	100.00	114.40	103.89
3 rows	67.51	107.17	48.58
5 rows	40.20	30.46	56.55

When the probable errors as determined by means of the analysis of variance are used for the efficiency test different results are obtained, since the error due to location in the field has been eliminated. The results from the analysis of variance, as given in the lower part of Table 109, show that certain of the larger plots have a higher percentage of efficiency than the smallest unit. For example, a single-row plot 30 feet in length has an efficiency percentage of 114.40 as compared with the standard 20-foot plot, which indicates that from the analysis of variance the larger plot would be more satisfactory.

The choice of the size of plot depends on a further factor. As stated earlier, on account of the variability of plots it has been found desirable to have more than one plot of each treatment, or, in other words, plots must be replicated. Therefore, in determining the size of plot the number of replications to be used must also be considered. Data from a blank test may also be used in determining the number of replications necessary by assuming different rows to be replicates of the same variety or treatment. For example, with the blank test of cotton we have 140 rows 100 feet long. This gives 700 20-foot units. One may study the effect of replications by assuming that the first 350 rows are sown to different varieties and that the second 350 rows are duplicates of the first. This will give one replication, that is, row 1 and row 351 will be considered as duplicates. Again, if one desires to study nine replications, he may assume that the first 70 rows are sown to 70 different varieties, and beginning with row 71 he would have the series replicated once. The second replication would begin with row 141, the third replication with row 211, and so on. Other combinations of replications, together with the different lengths and widths, may be studied, thus determining the effect of replications. This has been done with the data from the blank test of cotton, and the probable errors for plots of different sizes and number of replications are given in Table 110, page 417.

In Chapter XII it was shown that the probable error of the mean of any number of determinations equals $P.E._s/\sqrt{N}$. In Table 108 it was shown that the probable error of a single 20-foot plot is 12.56. From this we would expect the probable error of two single-row plots 20 feet long to be $12.56/\sqrt{2}$, or 8.88. Referring to Table 110 it is seen that the actual probable error is 9.01. This agrees rather closely with the result expected from theory. For the single 30-foot row in Table 108 we find the probable error to be 11.34. Then, for two such rows we would expect the probable error to be $11.34/\sqrt{2}$, or 8.02. The probable error actually obtained, as seen in Table 110, is 8.47.

The value of replications may be shown by the data in Table 110. The probable error for a 2-row plot 20 feet long is 10.35 per cent, while the probable error for two single 20-foot plots is 9.01. Again, the probable error for a single 4-row plot is 9.51, while for the same area made up of four scattered units (three replications) it is 7.55. For a 4-row plot 40 feet long the probable error is 7.29 per cent, while for the same area made up of four scattered units 40 feet long the probable error is 5.63 per cent.

TABLE 110

PROBABLE ERRORS IN PER CENT OBTAINED FROM REPLICATING PLOTS OF DIFFERENT SIZES

WIDTH OF PLOT	NUMBER OF REPLICATIONS	NUMBER OF PLOTS	LENGTH OF PLOT IN FEET				
			20	25	30	35	40
1 Row	0	1	12.56	11.97	11.34	10.68	10.09
	1	2	9.01	8.73	8.47	7.92	7.42
	2	3	7.70	7.96	6.70	6.57	6.37
	3	4	7.55	6.51	5.94	5.71	5.63
	4	5	5.21	4.49	4.86	4.95	4.36
	6	7	5.19	4.04	3.77	3.95	3.45
2 Rows	0	1	10.35	9.74	9.61	9.05	8.40
	1	2	7.56	7.02	7.09	6.56	6.08
	2	3	6.10	6.06	5.87	5.79	5.48
	3	4	6.42	5.12	5.48	4.76	4.86
	4	5	3.96	3.44	3.68	3.34	3.48
	6	7	4.50	3.24	3.20	3.25	2.93
3 Rows	0	1	9.37	9.02	8.74	7.85	7.54
	1	2	6.92	5.90	6.76	6.18	5.91
	2	3	5.34	5.79	5.48	4.65	5.29
	3	4	5.52	3.79	5.10	4.93	4.27
	4	5	3.36	3.30	3.36	3.29	2.60
	6	7	3.83	2.64	2.53	2.75	2.38
4 Rows	0	1	9.51	8.93	8.67	7.96	7.29
	1	2	6.74	5.69	6.48	6.29	5.58
	2	3	5.49	5.74	5.56	5.31	5.54
	3	4	5.51	3.95	5.63	4.62	4.24
	4	5	3.42	2.74	3.16	2.97	2.79
	6	7	4.23	2.91	2.58	2.96	2.39

Further comparisons may also be made between the larger plots and the same area made up of scattered units of smaller plots. For example, the probable error of one 40-foot plot is 10.09 per cent, while for the same area made up from two 20-foot plots it is 9.01 per cent. Again, a 3-row plot 40 feet long has 120 feet of row and the probable error is 7.54 per cent, while the probable error for four single 20-foot plots having a total of only 80 feet of row is 7.55 per cent.

From these and other comparisons that may be made it is evident that greater accuracy may be obtained by replicating smaller plots than by using larger single plots. Smaller units will require a little more land for borders or boundaries between series of plots, but in general the smaller plots with more replications are more desirable.

From these results and considering the other factors involved, that is the preparation of the soil and the handling of the plots in the field, the number of plants to be grown on a plot, and the like, one would select the size of plot that will be most convenient for the type of experiment being conducted and at the same time give a probable error low enough to enable him to measure the differences he needs to determine. The smaller the differences that one must measure the lower must be the error and consequently the greater the number of replications.

This may be made clear by using the data in Table 110. The probable error of the mean of a 20-foot plot replicated three times, giving four plots in all, is 7.55 per cent. If one is interested in comparing the difference between two varieties or treatments for this length of row and number of replications, the probable error of the difference will be $7.55\sqrt{2}$, or 10.68 per cent. Taking 3.2 as the ratio giving odds of approximately 31 to 1, it is necessary then that the percentage difference shall be 10.68×3.2, or 34.18 per cent. In other words, the smallest significant difference that can be measured with a single 20-foot plot replicated three times is 34.18 per cent. That is, if we were comparing two varieties of cotton, one of which yields 100 catties of seed cotton per mow, the other variety must yield more than 134 or less than 66 catties per mow before we can be sure they are significantly different in their yielding ability. This is on the basis of four single 20-foot plots. Again, if one were using single 20-foot plots with six replications, or seven plots in all, the probable error of the mean of the seven plots is 5.19 per cent, and the lowest difference which one can measure and have the difference significant is $5.19\sqrt{2} \times 3.2$, or 23.49 per cent. This illustrates how one may use the results of a blank test to determine how small a difference for plots of various sizes it is possible to measure and have the result lead to significant odds.

One may continue this comparison by considering a single-row plot and a 3-row plot. It was shown that with a single-row plot replicated six times the smallest difference that could be measured was 23.49 per cent. With a 3-row plot replicated six times it is possible to measure a difference of $3.83\sqrt{2} \times 3.2$, or 17.34 per cent. However, while a smaller difference can be measured, the 3-row plot requires three times as much land as a single-row plot, and the reduction in the percentage difference that can be measured is only 23.49–17.34, or 6.15 per cent.

It will be of interest to learn what difference can be measured with the same amount of land as was used for the 3-row plots replicated six times by using single-row plots rather than 3-row plots. For the 3-row plots replicated six times we have a total of 3×7, or 21 rows. The probable error for the mean of 21 single 20-foot rows would be expected to be $12.56/\sqrt{21}$, or 2.74 per cent. The smallest significant difference, therefore, that can be measured with 21 single 20-foot plots is $2.74\sqrt{2} \times 3.2$, or 12.38 per cent. It is evident that it is possible to measure smaller differences with the same amount of land planted to single-row plots, thus permitting more replications. This is true provided no other factors, such as competition, which will be discussed later, affect the results.

The results of these various comparisons illustrate how blank tests may be used in determining the size and shape of plots and the number of replications, and on the basis of these and similar results one may determine the plot arrangement that will lead to the desired degree of accuracy. It must be kept in mind at all times that while the analysis of a blank test may indicate the number of plots or plot arrangement most suitable for an experiment, it may not always be possible to follow in practice the arrangement that calculations have indicated is the best. That is, owing to shortage of land or labor, it may not be possible to handle as many plots as the calculations have indicated are the most desirable for a particular kind of experiment.

As indicated earlier, one may conduct blank tests for different crops and for different localities. It may be pointed out, however, that it is not necessary to conduct such tests with every kind of crop or for each particular set of environmental conditions. For example, the results obtained from a blank test of cotton will be applicable to other crops requiring a similar amount of space per plant. The same is true with results obtained from a blank test of wheat, since such results would be applicable to other small grain plots.

Competition. During recent years many studies have been made to determine border effect and the effect of competition between plots. In experimental work it is often desired to study the yielding ability of different varieties of crops or the effect of different treatments, and in many instances varieties differing considerably in their growth habits, time of maturity, and the like, or plots receiving different amounts of fertilizer, are tested side by side. In such cases it is important to know what effect may be caused by the unequal growth on two adjacent plots. That is, what will be the effect if a vigorous-growing variety is in a plot next to a poor variety. Will the poorer variety be adversely affected so that its yield is actually lower than it would be if there were no competition, and will the vigorous-growing variety yield more since it is growing next to a poor variety than it would if it were growing next to a standard variety or another plot sown to the same variety? This question of competition is very important since it has an influence on determining the size and shape of plots to be used.

Various methods have been used to measure the effect of competition. Stadler gives results obtained by Kiesselbach, which are presented here as Table 111, page 421. The method used to determine competition was to grow different varieties in single alternating rows and then grow the same varieties in alternating 5-row blocks to see whether the relative yields under the two conditions were the same. The crops were wheat and oats, and it will be noted that there is a difference in several of the comparisons. For example, for the wheat test in 1913, in alternating rows Big Frame yielded 107 per cent compared with 100 per cent for Turkey, while it yielded only 97 per cent in alternating 5-row blocks. The comparison between the Burt and Kherson oats in 1913 shows that Burt outyielded Kherson by 30 per cent in alternating rows, while it yielded only 12 per cent more in alternating 5-row blocks.

TABLE 111

THE EFFECT OF COMPETITION, EXPRESSED IN PER CENT, BETWEEN THE YIELDS OF VARIETIES GROWN IN ALTERNATING ROWS AND IN ALTERNATING 5-ROW BLOCKS

WHEAT AVERAGE YIELD OF 50 PLOTS			OATS AVERAGE YIELD OF 50 PLOTS		
YEAR AND VARIETY	ALTERNATING SINGLE ROWS %	ALTERNATING 5-ROW BLOCKS* %	YEAR AND VARIETY	ALTERNATING SINGLE ROWS %	ALTERNATING 5-ROW BLOCKS* %
1913			1913		
Turkey	100	100	Kherson	100	100
Big Frame	107	97	Burt	130	112
1914			1914		
Turkey	100	100	Kherson	100	100
Big Frame	85	97	Burt	139	101
1913			1913		
Turkey	100	100	Kherson	100	100
Nebraska No. 28	107	107	Swedish Select	82	77
1914			1914		
Turkey	100	100	Kherson	100	100
Nebraska No. 28	63	85	Swedish Select	89	93

*Yield based on 3 inner rows of 5-row plots in 1914

Stadler presents a number of examples from his own studies, and one comparison is given here as Table 112.

TABLE 112

THE EFFECT OF COMPETITION, EXPRESSED IN PER CENT, BETWEEN THE YIELDS OF VARIETIES AS OBTAINED FROM INTERIOR ROWS AND FROM COMPETING BORDER ROWS

VARIETY	YIELD IN INTERIOR ROWS		YIELD IN COMPETING BORDER ROWS	
	BUSHELS	RELATIVE	BUSHELS	RELATIVE
Leap's Prolific	14.9	91	9.9	53
Michigan Wonder No. 116	16.4	100	18.8 21.7	100 100
Poole Selection	15.3	93	11.5	53

This table shows the relative yields when varieties of wheat are grown side by side and comparisons made from the inner rows and from the border rows of 5-row blocks. The yield of Leap's Prolific is 91 per cent of the yield of Michigan Wonder on the basis of the yield of the interior rows, while when the border rows are compared Leap's Prolific yields only 53 per cent as much as Michigan Wonder.

These examples illustrate the effect of competition and show one method for measuring this effect. There are other ways of determining whether there is any disturbing effect of competition. For example, if one were comparing varieties of grain in 3-row blocks, if competition is present he would expect the yields of the middle rows to be less variable than the yields of the outside rows. It is possible, therefore, to compare the coefficients of variability for the middle rows with the coefficients of variability for the outside rows. If competition is having any serious effect we would expect the coefficients of variability for the middle rows to be less than for either of the outside rows. This is especially true if the same varieties do not grow next to each other in all of the replicated blocks.

Another method which may be used, when comparisons are made in 3-row blocks, to determine whether competition is having any serious effect on the yield or the comparisons in yield between the different varieties, is to determine the yield from all three rows of each block and then determine the yield from the middle row only. The varieties may be ranked in accordance with these yields and the correlation between the ranks obtained. If the correlation is high it indicates that there is no serious disturbing effect so far as yield is concerned. This method will not determine the amount of competition, but it will show whether competition is having any serious effect on the yield.

Two other methods for determining the effect of competition, which are similar to each other, are those used by Stringfield and Shen. We will apply these methods to the data in Table 113.

TABLE 113

APPLICATION OF THE METHODS OF SHEN AND STRINGFIELD FOR DETERMINING THE EFFECT OF COMPETITION

VARIETY A			VARIETY B		
Row 1	Row 2	Row 3	Row 1	Row 2	Row 3
51	50	48	45	37	33
43	51	47	55	55	49
43	38	29	44	40	45
47	46	49	49	54	49
41	36	37	57	37	36
29	34	40	43	42	36
34	35	41	34	33	32
33	33	38	32	33	32
38	20	27	30	31	26
35	39	43	22	32	36

TABLE 113—*Continued*

THE DATA ANALYZED BY SHEN'S METHOD

VARIETY A		D_1	VARIETY B		D_2	D_1-D_2	D'	D'^2
YIELD OF Row 2	YIELD OF Row 3		YIELD OF Row 2	YIELD OF Row 1				
50	48	2	37	45	− 8	10	12	144
51	47	4	55	55	0	4	6	36
38	29	9	40	44	− 4	13	15	225
46	49	−3	54	49	5	− 8	− 6	36
36	37	−1	37	37	0	− 1	1	1
34	40	−6	42	43	− 1	− 5	− 3	9
35	41	−6	33	34	− 1	− 5	− 3	9
33	38	−5	33	32	1	− 6	− 4	16
20	27	−7	31	30	1	− 8	− 6	36
39	43	−4	32	22	10	−14	−12	144

$\Sigma = -20$ $\Sigma = 656$

$M = \frac{-20}{10} = -2.0$ $S.D. = \sqrt{\frac{656}{10}}$

$= 8.1$

$Z = \frac{-2.0}{8.1} = -.25$ Odds not significant

THE DATA ANALYZED BY STRINGFIELD'S METHOD

YIELD OF Row 3 VARIETY A	YIELD OF Row 1 VARIETY B	D_1	YIELD OF Row 2 VARIETY A	YIELD OF Row 2 VARIETY B	D_2	D_1-D_2
48	45	3	50	37	13	−10
47	55	− 8	51	55	− 4	− 4
29	44	−15	38	40	− 2	−13
49	49	0	46	54	− 8	8
37	37	0	36	37	− 1	1
40	43	− 3	34	42	− 8	5
41	34	7	35	33	2	5
38	32	6	33	33	0	6
27	30	− 3	20	31	−11	8
43	22	21	39	32	7	14

$\Sigma = 20$

$M = \frac{20}{10} = 2.0$

The odds will be the same as obtained by Shen's method above.

These data are the yields from 3-row blocks of wheat replicated nine times. The yields of the individual rows for the two varieties are given at the top of the table in the order in which the varieties were grown in the field. Row 3 of variety A grew next to row 1 of variety B.

The method used by Shen will be explained first. This consists of obtaining first the difference between the middle row and the outside row, or the row next to the variety which is being used in the comparison. This is done pair by pair, as illustrated in Table 113. The same method is followed for the second variety, that is, the difference between the middle row and the outside row growing next to the first variety is obtained, as illustrated in the table. This gives two series of differences. The difference between these is obtained by subtracting in turn the second series of differences from the first series of differences, having regard to the signs. The remainder of the analysis is made by determining 'Student's' Z value and obtaining the odds. That is, the mean difference and the standard deviation are obtained, from which it is possible to determine Z and the odds. If the effect of competition is important the odds will be high enough to indicate significance. If the odds are low it indicates that there is no serious effect from competition.

Stringfield's method, quoted by Shen, is applied to he same data and is illustrated in the lower part of Table 113. By this method the differences between the adjacent border rows are first obtained and then the differences between the middle rows of the two varieties being compared are obtained. The second series of differences are subtracted from the first series of differences, and this leads to the same numerical values for the differences between the two series as obtained by Shen's method, although in certain instances the signs will not be the same. 'Student's' method may be used for the analysis and the interpretation made in the same way.

These methods make it possible to determine for each variety how much competition is really present, and when a similar study for the same varieties is conducted for two or three years it is possible to determine whether certain varieties are more affected by competition than others.

Stadler has applied a different method to measure the effect of competition. In discussing this method he states: "The average yield of each border row for each variety was converted to the percentage of the average yield of the same variety in its interior rows. These yields of border rows in percentage will be referred to as 'relative border yields.' The relative border yield gives a rough indication of the effect of competition on the variety. When it is above 100, the variety yielded more in border rows (subject to competition) than in interior rows (protected from competition). When it is below 100, the border yield was less than the interior yield, in proportion.

"An approximate measure of the competition between each pair of adjacent varieties was obtained by dividing the higher relative border yield by the lower, in the case of their adjacent border rows, and subtracting 100 from the result. . . . It will be referred to, for convenience, as the coefficient of competition." The steps may be illustrated by data presented by Stadler and given here as Table 114.

TABLE 114

DATA USED TO ILLUSTRATE STADLER'S METHOD FOR DETERMINING THE COEFFICIENT OF COMPETITION

	FULTZ			MICHIGAN AMBER			MICHIGAN WONDER No. 211		
	Row 1	Row 2, 3, 4	Row 5	Row 1	Row 2, 3, 4	Row 5	Row 1	Row 2, 3, 4	Row 5
Average yield in bushels	10.8	12.2	13.1	13.3	14.9	14.5	19.8	18.1	19.4
Yield expressed in per cent	89		107	89		97	109		107

Stadler continues: "Now dividing the yields in border rows by the yields of the same varieties in interior rows, we obtain the relative border yields. . . . To determine the degree of competition between the varieties Fultz and Michigan Amber, we divide the larger relative border yield (107) by the smaller (89) and subtract

100, giving 20 per cent. Since in this case the relative border yield of the variety on the left is higher, the difference is given a minus sign. Similarly a value of +12 per cent is obtained for the competition between Michigan Amber and Michigan Wonder No. 211. These figures mean that the relative border yield of Fultz exceeded that of Michigan Amber by 20 per cent in their competing border rows, while that of Michigan Wonder exceeded that of Michigan Amber by 12 per cent. The relative yields of these varieties are obtained similarly,—in the first case by dividing 14.9 by 12.2 (+22%) and in the second case by dividing 18.1 by 14.9 (+21%)." These coefficients of competition were then correlated with the yield and other characters to determine the effect of competition. The results obtained by correlating the competition coefficients with yield are shown in Table 115.

TABLE 115

COEFFICIENTS OF CORRELATION OBTAINED BETWEEN COMPETITION AND YIELD FOR DIFFERENT CROPS

TEST	SEASON	MEAN COEFFICIENT OF COMPETITION	COEFFICIENT OF CORRELATION BETWEEN COMPETITION AND YIELD
Barley variety	1919	21.30	.442 ± .099
Oats variety	1919	27.67	.314 ± .117
Oats strain	1919	13.11	.316 ± .143
Wheat variety	1920	19.79	.582 ± .043
Wheat variety	1921	18.85	.294 ± .059
Wheat mixture	1921	14.28	.554 ± .078
Oats variety	1921	39.15	.484 ± .082

These results indicate that so far as Stadler's data are concerned there is a rather high correlation for most of the comparisons, but when one is testing different varieties to determine their relative yields the important question is whether competition has disturbed the yields to such an extent that the relative yielding ability is different when the effect of competition is removed from that when the effect of competition is not removed. For example, in Table

115 where the coefficient of competition for the oats variety test for 1921 is correlated with yield, the coefficient of correlation is .484±.082. This shows that competition is present, but when these varieties are ranked according to yield, using all the rows of the plot and then eliminating the border rows and the effect of competition, it is found that the correlation between ranks is very high. In fact it is .99±.002 and the first ten varieties are the same in each case whether the yield is based on all the rows of the plot or on the inner rows only.

Again, for the barley variety test where the correlation is shown to be .442±.099, the correlation between the ranks arranged according to the yield of all the rows of the plot and the yield of the inner rows only is .99±.003, and the first few highest yielding varieties are the same whether the yield is based on the inner rows or on all the rows of the plot. These data show that competition may be present even to a rather high degree and yet the ranking of the varieties may not be seriously disturbed.

The effect of competition will vary in accordance with environmental conditions and locality, and it is important to consider its effect in arranging planting plans and plot layouts. With certain types of experiments it is very important that every precaution be taken to eliminate any possible effect of competition. This is especially true with reference to cultural and soil fertility experiments, since in this type of experimental work the plots are to be located permanently, or at least for several years. For example, in soil studies one will plan to give certain plots a definite treatment which is to be continued over a period of years, and in such cases it is important that the plots be so arranged as to eliminate any border effect. This can be done by having the plots of a larger size and then cutting off a border or strip on the sides and ends of the plots. In such experiments as soil fertility studies it is important to have a border several feet wide between the treatment plots. This will require additional land but it will insure more accurate results. Similar arrangements should be made for plots that are to be used for rotation and cultural experiments.

For the study of the yielding capacity of varieties, as is necessary in connection with plant improvement work, if competition is present to a decided degree it may be overcome by planting blocks of three or more rows for each variety and harvesting only the center row or rows. This requires an added amount of land and it is possible that in many localities competition is not so serious but that its effect may be offset by grouping together those varieties that are similar in their growth habits, maturity, and the like. Thus one may have all of the early varieties planted in adjoining plots, and so on. It is likely that in many localities when the varieties under test are grouped in accordance with their habit of growth it will be possible to arrange the series in adjacent plots using even single-row plots, without encountering serious competition. It is generally true that plant-breeding investigations conducted in a particular locality are confined largely to the study of types similar in their general behavior. For this reason, especially for the preliminary tests, it will be practical to use single-row plots, and for the final elimination test to have blocks of three or more rows. The rows may be harvested separately and if competition is found to be serious the yields of the middle rows only may be used.

Check Plots. The use of check plots has been the cause of considerable discussion, especially in recent years. Opinions are varied as to the usefulness of check plots, and whether it is desirable to have many or few check plots in a plot layout.

From the very early experiments down to those of the present time, most plot layouts have included some sort of check plots to determine the effect of treatment, or to serve as a standard for the comparison of varieties. In some of the early investigations having to do with soil fertility studies, the check plots were given no treatment whatever. Naturally, over a period of years these check plots gradually decreased in yielding ability. This indicated that it was unfair to compare the effect of treatment with the check plot, since owing to the fact that the check plot was gradually decreasing in its yielding ability the effect of treatment was being exaggerated. It was decided that for such studies the check plot should receive sufficient soil treatment to maintain its fertility to the degree possessed at the beginning of the experiment. This has led to a change in the plan of the check plots in soil studies so that the check plots are no longer 'nothing' plots, that is plots receiving no fertilizer, but they are thought of as control plots where efforts are made to maintain the natural fertility.

For such studies as concern the plant breeder in comparing different strains, the check plots used are sown to a standard variety whose yielding performance and other characteristics are known, and these check plots are used as a basis for comparing the yielding ability and other characteristics of the new strains. There is a difference of opinion among investigators as to the number of checks that should be used in such studies. Some believe that the checks should be very frequent, and others believe that it is more desirable to have fewer checks and possibly to have more replications of the strains under investigation. With the newer methods of plot arrangement and the analysis of results, that is random arrangement and the analysis of variance, there is a tendency to have only a few, or no, check plots. In most instances it is desirable to have a standard variety to be used for comparison, and when following random arrangement this variety or check will be considered as one of the varieties under test and will be handled the same as the other varieties.

In experimental work where it seems desirable to use checks the best arrangement and frequency of the check plots may be determined from the blank test if such a test is available. In using the data from a blank test to determine the arrangement and frequency of checks, the checks are located at certain intervals, for example every third or every fifth plot, and the deviations of the yields of the other plots from the check yields are determined. The arrangement that gives the smallest deviations is considered to be the best, but it may not be the most practical from the standpoint of the experiment.

In determining these deviations different methods may be used. A simple method is to let the average of the two nearest check plots represent the yield for the plots between these two check plots. For example, if the check plots are arranged every third plot there will be two plots between each pair of checks. Suppose plots 1 and 4 are check plots and their yields are 114 and 126 catties per mow. Their average yield is 120 catties, and if plots 2 and 3 yield 125 and 130 catties, then the deviations of these plots from the average of the two check plots would be determined, giving deviations of 5 and 10 catties

Another method is to take the average of all check plots in the field and then obtain the deviations of the other plots from the average of these check plots. Still another method, and the one very frequently used, is that of obtaining the graded difference between the check plots. Thus, for the two check plots cited above, giving yields of 114 and 126 catties, we obtain the difference between the two checks and then calculate the check yields for the intervening plots on the assumption that there is a gradual change in the fertility of the soil from one check to another. Since the checks are located every third plot, we divide the difference, 12, by 3, giving 4, which is the assumed difference between the plots. Since the first check is lower in yield, it is assumed that as we proceed to the next check the soil improves by 4 units for each plot. Following this method the calculated yield for plot 2 is 114+4, or 118, and for plot 3 it is 118+4, or 122. If the first check plot had been the higher in yield, then the average difference would have been subtracted. This process may be put in a formula as follows:

$$\frac{2}{3}C_1+\frac{1}{3}C_2=\text{Calculated check for first value}$$

$$\frac{1}{3}C_1+\frac{2}{3}C_2=\text{Calculated check for second value}$$

Here C_1 and C_2 represent the yields for the check plots, and substituting the actual values from the example we have

$$\frac{2}{3}114+\frac{1}{3}126=118$$

$$\frac{1}{3}114+\frac{2}{3}126=122$$

The deviations between the calculated check plots and the yields of the variety plots are then determined.

Another method that may be used for comparing test plots with the calculated yield of the check plots is to use a combination of the graded method with the average of all the check plots. The different methods for obtaining calculated check yields may be represented by the following formulas.

METHOD	HOW OBTAINED
1	Average of two nearest checks, $\frac{C_1+C_2}{2}$ = Calculated check
2	Average of all check plots, $\frac{C_1+C_2+C_3+\ldots C_n}{N}$ = Calculated check
3	Graded method Check every fifth plot, $\frac{4}{5}C_1+\frac{1}{5}C_2$ = Calculated check for first value $\frac{3}{5}C_1+\frac{2}{5}C_2$ = Calculated check for second value, etc For check plots located at other distances, such as every tenth the formula is $\frac{9}{10}C_1+\frac{1}{10}C_2$ = Calculated check for first value, etc.
4	Graded method plus the mean of all checks $\frac{1}{2}\left(C_M+\frac{4}{5}C_1+\frac{1}{5}C_2\right)$ = Calculated check for first value, etc.
	In methods 3 and 4, C_1 and C_2 represent the actual yields of any two consecutive check plots, and C_M in method 4 represents the mean of all the check plots.

When using data from a blank test in which all of the plots are sown to the same variety to determine which arrangement of checks, that is a check every third, fourth, or fifth plot, gives the smallest deviations, one can select the arrangement of checks most suitable for the type of experiment to be conducted. Naturally, there are other considerations that will be important in determining the check arrangement to be followed. For example, one may find that the smallest deviations are obtained with every other plot as a check, but this means that half of the land would be given up to check plots. Such frequency of check plots may not be feasible nor even necessary to measure the differences obtained from the experiment.

The results from an analysis of a blank test of oats, where the check has been placed every third, fifth, and tenth plot, and the results compared by three methods of calculating the theoretical check yields, are given in Table 116.

TABLE 116

THE MEANS OF THE DEVIATIONS OBTAINED FROM DIFFERENT METHODS OF ARRANGING CHECKS. THE DATA, IN BUSHELS PER ACRE, ARE FROM A BLANK TEST OF OATS

CHECK EVERY TENTH PLOT	1921	1922	1923
Compared by method 2	1.71 ± .11	1.58 ± .11	2.66 ± .14
Compared by method 3	1.81 ± .13	1.52 ± .12	2.17 ± .12
Compared by method 4	1.79 ± .12	1.53 ± .12	2.07 ± .13
CHECK EVERY FIFTH PLOT			
Compared by method 2	1.90 ± .13	1.57 ± .12	2.70 ± .16
Compared by method 3	1.84 ± .13	1.58 ± .12	1.90 ± .12
Compared by method 4	1.84 ± .12	1.52 ± .12	2.10 ± .15
CHECK EVERY THIRD PLOT			
Compared by method 2	1.32 ± .09	1.56 ± .13	2.47 ± .18
Compared by method 3	1.18 ± .08	1.87 ± .17	1.76 ± .13
Compared by method 4	1.07 ± .09	1.67 ± .16	1.80 ± .14

The test was conducted for three years and three different methods have been used to determine the deviations of the test plots from the calculated check yields. One point to be observed is that in general there is little difference in the deviations obtained from the three different methods of using checks. This is especially true for the results obtained from checks every fifth and every tenth plot for the years 1921 and 1922. The data for 1923 show some variation from the results for the first two years. When a check plot is placed every third plot there is a tendency for the deviations to become smaller, especially for the years 1921 and 1923. On the other hand, for the year 1922 the deviation as measured by the graded method is higher than in cases where the check is located every fifth or tenth plot. These data show that for rod-row plots there is little difference between checks every fifth or every tenth plot. The deviations are somewhat smaller when the checks are every third plot.

Another illustration of the use of check plots arranged by different systems is shown by the data presented by Parker and Batchelor, given here as Table 117.

TABLE 117

RESULTS OBTAINED FROM DIFFERENT METHODS OF ARRANGING CHECKS AND OF CALCULATING THEORETICAL CHECK YIELDS

FREQUENCY OF CHECKS	METHOD OF CALCULATING THEORETICAL CHECK YIELDS	COEFFICIENT OF VARIATION OF TEST PLOTS, IN PER CENT		CORRELATION COEFFICIENT BETWEEN ACTUAL YIELD OF TEST PLOTS AND THEIR THEORETICAL CHECK YIELD OR PLOT VALUE
		BEFORE ADJUSTMENT	AFTER ADJUSTMENT	
No checks		13.15		
Check every third	1. Mean of checks	12.79	12.79	—
	2. Nearest check	12.79	10.40	.688 ± .042
	3. 2/3 C_1+1/3 C_2	12.79	9.08	.715 ± .040
	4. 1/2 (C_M+2/3 C_1+1/3 C_2)	12.79	9.46	.747 ± .035
Check every fifth or sixth	1. Mean of checks	13.26	13.26	—
	2. Nearest check	13.26	11.44	.599 ± .048
	3. 4/5 C_1+1/5 C_2	13.26	9.53	.590 ± .048
	4. 1/2 (C_M+4/5 C_1+1/5 C_2)	13.26	10.52	.599 ± .048
Check every seventh	1. Mean of checks	12.88	12.88	—
	2. Nearest check	12.88	11.59	.586 ± .048
	3. 6/7 C_1+1/7 C_2	12.88	10.06	.641 ± .042
	4. 1/2 (C_M+6/7 C_1+1/7 C_2)	12.88	9.51	.643 ± .042

In this test trees have been used as the crop studied. The four methods for using the check plots are: the mean of all checks, the nearest check, the graded method, and a combination of the graded method and the mean of all checks. The coefficient of variability has been used to measure the effect of the different arrangements of checks. It is to be noted that the lowest coefficient of variability is given by the graded method in two out of the three comparisons.

In addition to the coefficient of variability the correlation is given, showing the relation between the actual yields of the test plots and their theoretical check yields. It is seen that the correlation is high and significant in all cases, and in general the correlation is slightly higher for the graded method and the combination of the graded method and the mean of all checks than it is when the nearest check is used. This is important, since there has been some question as to the reliability of check yields calculated on the basis of the graded method.

The data from the blank test of cotton have also been used to determine the best arrangement of checks, and the results of this study are presented in Table 118, page 436. The data in this table are obtained from plots 30 feet in length and the checks are placed every third and fifth plot. The averages of the deviations for the various types of plots and for the different number of replications are given. In general the average of the deviations is larger for checks placed every fifth plot than for checks every third plot. The average for all tests shows an average difference of .099. This difference is not large and for the types of plots used here it would be better to have more replications with the checks every fifth plot. That is, if land is limited it is better to utilize the land for more replications rather than to have fewer replications with more frequent checks.

Richey has suggested another system for handling check plots and a method for adjusting yields on the basis of a moving average. The method that he suggests for the use of checks is that for one series one of the strains in the test will be planted in alternating check plots and all of the other strains compared with this one. For a second series another one of the strains in the test will be

TABLE 118

MEANS OF THE DEVIATIONS OBTAINED FOR PLOTS OF VARIOUS SIZES AND NUMBER OF REPLICATIONS FROM DIFFERENT ARRANGEMENTS OF CHECK PLOTS

NUMBER OF REPLICATIONS	NUMBER OF ROWS IN ONE PLOT	FREQUENCY OF CHECK	MEAN OF THE DEVIATIONS
0	1	Every third	1.951 ± .102
		Every fifth	2.160 ± .094
	2	Every third	1.473 ± .105
		Every fifth	1.707 ± .097
	5	Every third	1.305 ± .146
		Every fifth	1.662 ± .172
1	1	Every third	1.391 ± .072
		Every fifth	1.430 ± .071
	2	Every third	1.033 ± .076
		Every fifth	.969 ± .068
	5	Every third	1.178 ± .102
		Every fifth	1.475 ± .171
2	1	Every third	1.133 ± .063
		Every fifth	1.089 ± .056
	2	Every third	.850 ± .081
		Every fifth	.937 ± .096
	5	Every third	.640 ± .075
		Every fifth	.800 ± .123
3	1	Every third	.948 ± .066
		Every fifth	.954 ± .071
	2	Every third	.737 ± .073
		Every fifth	.650 ± .066
	5	Every third	.861 ± .052
		Every fifth	.950 ± .102
4	1	Every third	.958 ± .081
		Every fifth	1.025 ± .063
	2	Every third	.733 ± .079
		Every fifth	.880 ± .086
	5	Every third	.757 ± .112
		Every fifth	.827 ± .118
6	1	Every third	.729 ± .069
		Every fifth	.857 ± .055
	2	Every third	.479 ± .068
		Every fifth	.597 ± .072
	5	Every third	.412 ± .103
		Every fifth	.385 ± .090

planted in alternating check plots and all of the other strains compared with it, and so on until each one of the strains has served as a check for one of the series. This arrangement is illustrated in Table 119 with data presented by Richey.

TABLE 119

PLANTING ARRANGEMENT FOR METHOD OF ADJUSTING YIELDS PRESENTED BY RICHEY, WITH RESULTS OBTAINED FROM THE FIRST SERIES

Row Number	Strain Numbers in Series					Results from First Series. Actual Yields in Pounds on Basis of Corrected Stand	
	1	2	3	4	5	Individual Rows	2-Alternate-Row Plots
1	10	10	10	10	10	9.8	
2	1	2	3	4	5	14.4	21.2
3	1	1	1	1	1	11.4	27.0
4	1	2	3	4	5	12.6	23.8
5	2	2	2	2	2	12.4	28.0
6	1	2	3	4	5	15.4	26.5
7	3	3	3	3	3	14.1	25.6
8	1	2	3	4	5	10.2	24.0
9	11	11	11	11	11	9.9	20.8
10	1	2	3	4	5	10.6	20.5

In the first series strain number 1 is used as a check for comparing the strains. For example, strain number 2 is planted in row 5, and the alternate rows 4 and 6 are planted with strain number 1 to serve as a check for comparing strain number 2. Strain number 3 is planted in row 7 and the check strain number 1 is planted in the alternate rows 6 and 8, and so on for the other strains. Thus, for that portion of the data presented here, strains number 1, 2, 3, and 11 have adjacent check rows sown to strain number 1. For the second series strain number 2 is used as a check and is planted in alternate rows. In the same way strains number 3, 4, and 5 are used as the check in the third, fourth, and fifth series, respectively.

The actual and adjusted yields for a number of strains, as presented by Richey, are given here as Table 120.

TABLE 120

THE MEAN ACTUAL AND ADJUSTED YIELDS OF 11 STRAINS OF CORN

STRAIN NUMBER	AVERAGE OF 22 REPLICATIONS*		
	ACTUAL YIELDS		ADJUSTED YIELDS ON BASIS OF 2-ALTERNATE-ROW PLOTS
	POUNDS	BUSHELS	BUSHELS
1	12.7	73.48 ± 1.50	72.61 ± 1.21
2	11.1	64.22 ± 1.45	63.99 ± .98
3	12.5	72.32 ± 1.45	73.25 ± .98
4	12.5	72.32 ± 2.43	72.90 ± 1.56
5	11.2	64.80 ± 1.62	67.81 ± 1.04
6	10.2	59.01 ± 1.16	56.35 ± 1.04
7	11.5	66.54 ± 1.62	68.73 ± 1.27
8	11.4	65.96 ± 1.21	64.51 ± 1.04
9	11.5	66.54 ± 1.68	66.88 ± 1.10
10	10.3	59.59 ± 1.68	58.49 ± 1.21
11	10.4	60.17 ± 3.53	64.11 ± 2.72
AVERAGE OF THE PROBABLE ERRORS		1.757	1.286

*Only 10 replications of strain number 11.

The adjusted yields were obtained in the following manner. The mean yields obtained from all the plots of the strains were computed, and these mean yields expressed in pounds, and also in bushels per acre with their probable errors, are given in Table 120. Richey assumed that the index of the productivity of the soil between any two alternate rows would be given by the ratio of the sum of the yields of these two alternate rows to the average yield of all the rows of those particular strains that were grown in the alternate rows in question. For example, the adjusted yield for row 6 in Table 119 is obtained in the following way. The ratio of the sum of the yields of rows 5 and 7 to the sum of the average yields of all the plots of strains 2 and 3, which are the strains that occur in rows 5 and 7, is obtained. The yield of row 5, as given in Table 119, is 12.4 pounds, and for row 7 the yield is 14.1 pounds. The sum of these yields is 26.5 pounds. The average yield of strain number

2, as given in Table 120, is 11.1 pounds, and for strain number 3 the average yield is 12.5 pounds. The sum of these average yields is 23.6 pounds. The ratio of the total yield of rows 5 and 7, 26.5 pounds, to the sum of the average yields of strains 2 and 3, 23.6 pounds, is 1.122, as obtained by Richey. The yield of row 6, 15.4 pounds, is divided by this ratio, giving the adjusted yield, which has been converted to bushels per acre. The yield of each row was adjusted in a similar manner and the average yield in bushels for each strain was determined. These average adjusted yields in bushels per acre, with their probable errors, are given in Table 120.

It is seen that there is a reduction in the probable errors in the case of the adjusted yields. The average probable error for the actual yields is 1.757 and for the adjusted yields it is 1.286, giving a reduction of .471 bushels. This indicates that a gain has been made by means of this method of adjustment, and smaller probable errors are available for making comparisons between the different strains. This method of using a moving average may be extended so that adjustments may be made on the basis of three or more adjacent rows.

Richey has also suggested the use of regression in determining adjusted yields. Hayes has followed this idea and has adapted the method to some of his data to obtain calculated yields based on the principle of regression analysis. By this method as adapted by Hayes the yield for each plot of a particular variety or treatment from the regular experiment is expressed as a percentage, considering the mean of each variety or treatment as 100 per cent. The correlation may therefore be determined between adjacent plots and between plots separated by one or more plots. If the series is arranged so that the check plots occur every fifth plot, then it would be necessary to determine the correlation only for adjacent plots and those separated by one, two, and three plots. The following data from Hayes, giving results obtained from spring wheat experiments, will illustrate the application of this method. These data have been arrived at in the following manner. If a variety in the first of the replicated plots yields 24 bushels and the mean yield for all replications of that variety is 25 bushels, its yield is expressed in per cent by dividing 24 by 25 and multiplying by

100, giving 96 per cent. The percentage yield of each row or plot in turn was obtained in a similar way and these percentages used for the correlations as presented here.

CORRELATION IN PERCENTAGE YIELDING ABILITY IN NEARBY PLOTS OF SPRING WHEAT, 1924

	CORRELATION OF	CORRELATION COEFFICIENT	REGRESSION EQUATION
Spring wheat rod rows	Adjacent plots	$.618 \pm .023$	$Y = 37.85 + .6205\ X$
	Separated by 1	$.518 \pm .028$	$Y = 47.99 + .5191\ X$
	Separated by 2	$.454 \pm .030$	$Y = 54.56 + .4546\ X$
	Separated by 3	$.383 \pm .034$	$Y = 61.30 + .3873\ X$
	Separated by 4	$.449 \pm .034$	$Y = 56.72 + .4350\ X$

The above correlation and regression coefficients were obtained in the usual way, and Hayes applied the regression coefficient to his data as follows:

PLOT NO.	ACTUAL YIELD	PERCENTAGE YIELD
Check A	25.1	93
8	27.5	
9	20.3	
10	28.3	
11	24.2	
12	25.7	
Check B	32.2	120

For the comparison of plot 8 we obtain the calculated yield, y, first on the basis of the regression for adjacent plots and then for plots separated by four. That is, the yield of plot 8 is calculated on the basis of the yield of check A and also on the basis of the yield of check B. The yield of each check is expressed as a percentage by dividing the actual yield by the average yield of all check plots. Using the regression coefficients for adjacent plots and for plots separated by four, we have

$$\text{For } x = 93 \quad y = 37.85 + .6205\ x = 95.56$$

$$\text{For } x = 120 \quad y = 56.72 + .4350\ x = 108.92$$

These two results are averaged, giving 102.24 per cent. The corrected yield is obtained by dividing the actual yield of each plot by the average calculated yield in per cent and multiplying by 100. For example, for plot [illegible] the actual yield, 27.5, divided by 102.24 per cent and multiplied by 100, gives 26.9 as the corrected yield. The same method would be used to obtain the corrected yields for all of the plots.

This method of adjusting yields will effect some reduction in the error of the experiment if the correlation is high. Richey states that when the correlations are less than .6 the adjustment will have little effect on reducing the variability, or the experimental error. The method is therefore not recommended unless the correlation between adjacent plots is very high, and in general it is found that when the method is used there is little change in the relative standing of the varieties.

The results of these different studies indicate that in general the deviations are less when the checks are placed close together. For many kinds of experiments it is important to use checks, but although the deviations are smallest when the checks are placed at frequent intervals it may not be necessary to locate the checks as close together as indicated by the preliminary data, since this will require a considerable area of land for the checks. With experiments in crop improvement where one is testing new varieties or strains, the check plots are of value for more than yield comparisons. For instance, for such problems as winter hardiness, disease resistance, stiffness of straw, and the like, we would use as a check variety one that is superior in these characteristics, so that in judging the value of the new strains the performance of the variety in the check plots is of great importance. The method of interpreting results on the basis of check plots will be discussed later.

Value of Uniformity Trials. It is often recommended that for a field which is to be used for cultural experiments or studies in crop rotation and soil fertility a uniformity study be conducted. By a uniformity study is meant that the field will be laid out in plots of the size and shape that are to be used in the actual experiments, and the entire field will be sown to the same crop. The field should be handled in as uniform a manner as possible and the study continued for several seasons. The plots will be harvested and the data used to determine the variability of the different plots. Some have criticized the value of uniformity trials on the basis that the yields of the plots may not be relatively the same from crop to crop, or that the plots will not react in the same manner after the experiment has been started. The data presented earlier in this chapter, illustrating the permanence of plot variability in the majority of cases, should answer the first objection, and at the present time the second objection is merely a supposition since not enough data are available to give accurate information on this point.

Fisher has given a method by which the data from uniformity trials may be analyzed and the results used for the interpretation of the data obtained from later experiments. This method had previously been applied by Sanders in his studies on the value of uniformity trials. In addition to the analysis of the variance this method considers the covariance, which is the mean product of the deviations of two variates, the deviations being measured from the respective means. The method may be illustrated with unpublished data furnished by R. J. Borden, of the Hawaiian Sugar Planters' Association, giving the yields obtained from the same plots for three different crops in uniformity trials of sugar cane. Sixteen plots have been used in this illustration, but the method may be applied to any number of plots. The yields for 1931 and 1933, expressed in per cent, are given in Table 121.

TABLE 121

THE APPLICATION OF THE ANALYSIS OF VARIANCE TO DATA FROM UNIFORMITY TRIALS, PRELIMINARY TO THE ANALYSIS OF COVARIANCE

	PRELIMINARY TEST				TOTAL
	113	113	91	102	419
	102	109	105	102	418
	96	99	94	101	390
	98	99	94	82	373
TOTAL	409	420	384	387	1600

GENERAL MEAN = 100

TABLE 121—*Continued*

	EXPERIMENTAL TEST				TOTAL
	103	110	92	113	418
	94	103	101	92	390
	103	99	94	99	395
	105	104	94	94	397
TOTAL	405	416	381	398	1600

GENERAL MEAN = 100

ANALYSIS OF VARIANCE

PRELIMINARY TEST

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE	STANDARD ERROR
Rows	3	378.5		
Columns	3	226.5		
Error	9	351.0	39.0000	6.24
TOTAL	15	956.0		

EXPERIMENTAL TEST

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE	STANDARD ERROR
Rows	3	114.5		
Columns	3	161.5		
Error	9	336.0	37.3333	6.11
TOTAL	15	612.0		

For convenience in the analysis these plots have been arranged in the form of a Latin square, but other methods of arrangement may be followed. We may consider the test for 1931 as the preliminary test and the results for 1933 as the results from the actual experiment. Applying the methods for the analysis of variance we have the results as given in the lower part of Table 121. The analysis shows that the standard error per plot for the preliminary test is 6.24 and for the experimental test it is 6.11.

We may now proceed to determine the covariance between the yields of the preliminary test and the experimental test in the following way. The values and the deviations in the preliminary test will be designated x and those in the experimental test y, and we will consider first the results from the rows. The mean of the row totals in the preliminary test is 400 and the total yield of the first row for the preliminary test is 419. Therefore the x deviation is 19. For the next row it is 18; for the third row, which gives a total yield of 390, it is −10; and for the fourth row it is −27. The y deviations for rows from the experimental test are determined from the mean of the row totals in the experimental test in the same manner. The squares of the x and y values and the products of x and y are obtained as illustrated in Table 122. The values for the columns for both the preliminary and experimental tests are obtained in a similar manner and also appear in Table 122, pages 445, 446. These values are summed and divided by the number of contributing units, giving the values to be used in the analysis of variance.

In addition to the results for rows and columns, it is necessary to have similar values for the individual plots, or the total. These are obtained by taking each plot in one test and determining its deviation from the mean of all the plots for that test. For example, the mean of the preliminary test is 100 and the first plot in that test yields 113, so the x deviation is 13. For the corresponding y plot in the experimental test the yield is 103 and the mean of the experimental test is also 100, so the deviation for y is 3. The other values for x and y are determined and together with their squares and products are shown in Table 122. These values are brought together in the summary. Subtracting the sums for rows and columns from the sum for the total we obtain the residue, or the amount due to error.

Table 122

Illustrating the Method of Determining Covariance and the Effect of Regression on Adjusted Yields

	x	y	x^2	y^2	xy
Rows	19	18	361	324	342
	18	-10	324	100	-180
	-10	- 5	100	25	50
	-27	- 3	729	9	81
			Σ = 1514	458	293
			Mean = 378.50[1]	114.50	73.25
Columns	9	5	81	25	45
	20	16	400	256	320
	-16	-19	256	361	304
	-13	- 2	169	4	26
			Σ = 906	646	695
			Mean = 226.50[2]	161.50	173.75
Totals	13	3	169	9	39
	13	10	169	100	130
	- 9	- 8	81	64	72
	2	13	4	169	26
	2	- 6	4	36	- 12
	9	3	81	9	27
	5	1	25	1	5
	2	- 8	4	64	- 16
	- 4	3	16	9	- 12
	- 1	- 1	1	1	1
	- 6	- 6	36	36	36
	1	- 1	1	1	- 1
	- 2	5	4	25	- 10
	- 1	4	1	16	- 4
	- 6	- 6	36	36	36
	-18	- 6	324	36	108
			Σ = 956	612	425

---

[1] Dividing by 4, the number of columns contributing to the individual deviations

[2] Dividing by 4, the number of rows contributing to the individual deviations

TABLE 122—*Continued*

SUMS OF SQUARES AND PRODUCTS

	DEGREES OF FREEDOM	x^2	xy	y^2
Rows	3	378.50	73.25	114.50
Columns	3	226.50	173.75	161.50
Error	9	351.00	178.00	336.00
TOTAL	15	956.00	425.00	612.00

ANALYSIS OF VARIANCE OF ADJUSTED YIELDS

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE	STANDARD ERROR
Rows	3	137.5175		
Columns	3	43.5391		
Error	8	245.7322	30.7165	5.54
TOTAL	14	426.7888		

From the values for x^2, y^2, and xy due to error we may determine the correlation coefficient, which will show the correlation between the yields of the plots for the different years. Having obtained the correlation coefficient we may then determine the regression coefficient. Dividing the values for x^2 and y^2 by the degrees of freedom, 9, and extracting the square root we obtain the standard deviations for x and y, which are 6.24 and 6.11, respectively. The correlation coefficient may then be determined from

$$r=\frac{\frac{\Sigma xy}{N}}{\sigma_x\,\sigma_y}$$

In this case N is the degrees of freedom, 9. Substituting the values for xy and for σ_x and σ_y we have

$$r=\frac{\frac{178.00}{9}}{(6.24)\,(6.11)}=.519$$

Substituting the necessary values in the regression equation for y on x, as given in Chapter VI,

$$y = r\frac{\sigma_y}{\sigma_x}\,x,$$

we have

$$y = .519\,\frac{6.11}{6.24}\,x, \text{ or}$$
$$y = .508\,x$$

Since for the purposes of the analysis we are interested only in the regression of y on x, this may be obtained directly, letting b equal the regression coefficient, or

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

Substituting the necessary values we have

$$b = \frac{178.00}{351.00} = .507$$

The slight difference in the values obtained for the regression is due to the handling of the decimals.

We may now use this regression coefficient, .507, to correct the values for rows, columns, and total, and thus obtain a new standard error of the experiment after the effect of regression has been eliminated. Since b is the regression coefficient the comparisons of adjusted yields are obtained from comparisons of $y - bx$. Now

$$(y - bx)^2 = b^2x^2 - 2bxy + y^2$$

and to obtain the sum of the squares for any line from the values given in Table 122 under the title 'Sums of Squares and Products' the entries in the table are multiplied by b^2, $-2b$, and unity (the coefficients of x^2, xy, and y^2) and the products summed.

Multiplying the values of x^2, xy, and y^2 for rows, columns, and total by the values for b^2 (.257049), $-2b$ (-1.014), and 1, respectively, we have the results for the analysis of variance after the effect of regression has been eliminated, as given in the lower part of Table 122. Thus for rows we have

$$\begin{aligned} 378.50 \times .257049 &= 97.2930 \\ 73.25 \times -1.014 &= -74.2755 \\ 114.50 \times 1 &= 114.5000 \end{aligned}$$

The sum of these three products is 137.5175. The values for the columns and total are obtained in a similar manner. The degrees of freedom for rows and columns remain the same, but one degree of freedom has been used in determining the regression and therefore the degrees of freedom for the total are 14 rather than 15. Adding the sums of squares and degrees of freedom for the rows and columns and subtracting from the total, we have the residue and degrees of freedom as given in Table 122. Dividing the residue, 245.7322, by the degrees of freedom, 8, we have 30.7165, and extracting the square root we have 5.54 as the standard error of the experiment after the effect of regression has been eliminated. This may be compared with the standard error, 6.11, obtained before the effect of regression was removed.

The same result will be obtained if the predicted or calculated yields for y are obtained from $y-bx$ and these calculated yields used for the analysis of variance in the usual way. The predicted yields for the individual plots have been obtained from $y-bx$ and are given in Table 123, page 449.

For the first y plot we proceed as follows. The first x plot in the preliminary series yields at the rate of 113 per cent of the mean yield, and the deviation from the mean is therefore 113–100, or 13. The adjusted yield for the first plot is obtained by substituting this value for x in the equation $y-bx$. The actual yield of the first y plot is 103, and we have $103-(.507\times 13)$, or 96.409, as the adjusted yield. For the next y plot, which yielded 110, we have $110-(.507\times 13)$, or 103.409. The other adjusted yields are obtained in a similar manner, and by applying the method of variance analysis we have the results as shown at the bottom of Table 123. These results are the same as given in Table 122, which were obtained by a shorter process.

It should be pointed out that in using the equation $y-bx$ it contains one value, b, which has been obtained from the data, therefore the degrees of freedom will be reduced from 15 to 14, and as b has been calculated from the values for error this degree of freedom will be taken from the degrees of freedom for error.

TABLE 123

APPLICATION OF THE ANALYSIS OF VARIANCE TO YIELDS ADJUSTED BY THE EQUATION $y - bx$, IN WHICH y IS THE ACTUAL YIELD

y	−	bx		
103	−	.507 × 13	=	96.409
110	−	.507 × 13	=	103.409
92	−	.507 × − 9	=	96.563
113	−	.507 × 2	=	111.986
94	−	.507 × 2	=	92.986
103	−	.507 × 9	=	98.437
101	−	.507 × 5	=	98.465
92	−	.507 × 2	=	90.986
103	−	.507 × − 4	=	105.028
99	−	.507 × − 1	=	99.507
94	−	.507 × − 6	=	97.042
99	−	.507 × 1	=	98.493
105	−	.507 × − 2	=	106.014
104	−	.507 × − 1	=	104.507
94	−	.507 × − 6	=	97.042
94	−	.507 × −18	=	103.126

ADJUSTED YIELDS USED FOR THE ANALYSIS OF VARIANCE

					TOTAL
	96.409	103.409	96.563	111.986	408.367
	92.986	98.437	98.465	90.986	380.874
	105.028	99.507	97.042	98.493	400.070
	106.014	104.507	97.042	103.126	410.689
TOTAL	400.437	405.860	389.112	404.591	

ANALYSIS OF VARIANCE OF ADJUSTED YIELDS

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE	STANDARD ERROR
Rows	3	137.5175		
Columns	3	43.5391		
Error	8	245.7322	30.7165	5.54
TOTAL	14	426.7888		

The foregoing illustrates the value of a preliminary trial in obtaining corrected yields for the plots and for determining a standard error for the experiment after eliminating the effect of regression. If the preliminary trial has been continued for several seasons or crops the results for each year may be analyzed separately, or it is possible to combine the results for several years, determine the average, and use this average yield in making the analysis. To illustrate how this may be done, the yields for the plots for the crop years 1929 and 1931 have been averaged and expressed in per cent, as shown in Table 124.

TABLE 124

APPLICATION OF THE ANALYSIS OF VARIANCE TO YIELDS ADJUSTED ON THE BASIS OF THE YIELDS OF PRECEDING CROPS

THE AVERAGE YIELDS OF THE CROPS FOR 1929 AND 1931 EXPRESSED IN PER CENT

					TOTAL
	106	110	95	108	420
	102	107	106	99	414
	97	97	95	99	388
	95	99	99	85	378
TOTAL	400	414	395	391	1600

EXPERIMENTAL TEST, 1933

					TOTAL
	103	110	92	113	418
	94	103	101	92	390
	103	99	94	99	395
	105	104	94	94	397
TOTAL	405	416	381	398	1600

TABLE 124—*Continued*

SUMS OF SQUARES AND PRODUCTS

	DEGREES OF FREEDOM	x^2	xy	y^2
Rows	3	306.00	86.50	114.50
Columns	3	75.50	84.25	161.50
Error	9	250.50	216.25	336.00
TOTAL	15	632.00	387.00	612.00

ANALYSIS OF VARIANCE OF ADJUSTED YIELDS

VARIATION DUE TO	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE OR VARIANCE	STANDARD ERROR
Rows	3	193.1003		
Columns	3	72.3146		
Error	8	149.3171	18.6646	4.32
TOTAL	14	414.7320		

With these values the calculations are carried out exactly as explained before. The values obtained for rows and columns are given in Table 124, as well as the results for the total. The regression, *b*, equals 216.25/250.50, or .863. Obtaining b^2 and $-2b$ the corrected values for the analysis of variance are determined and are given at the bottom of Table 124. It is seen that the standard error has been reduced to 4.32 by combining the yields of the two preliminary trials.

The predicted yields may be obtained on the basis of this regression value and the standard error determined by the usual method. It is possible, therefore, to determine the predicted yields for the various plots if the preliminary test has been conducted a sufficient length of time. In some cases there will probably be no reduction in the standard error as a result of preliminary trials, or, in other words, the correlation between the yields of the plots will be so low that there will be little gain effected by the regression. In many experiments, however, it is quite possible that there will be correlation between results for the different years and it may be well worth while to conduct uniformity trials.

It may be pointed out in this connection that even though uniformity trials may not result in a reduction of error, it is important when beginning work with new experimental fields to conduct uniformity trials at least for one or two crops, since it may be possible that parts of the field will be found unsuitable for experimental purposes, and this fact alone is worth the time and labor required for conducting the preliminary trials. If the results of the preliminary trial indicate decided variation in the plots, it may be important to continue the preliminary trials and then by the method of analysis just described determine the variance for the predicted or adjusted yields.*

Methods of Interpreting Results. There are various methods that may be used for interpreting the results obtained from field plots. The results are usually interpreted on the basis of probability, or by determining a standard error or probable error. This enables one to determine how much variation between results may be expected due to chance variation. The differences obtained from the experiments may then be compared with the differences expected from chance variation, and when the results from the experiments are enough larger than those expected from chance variation to lead to significant odds it is safe to conclude that there is a real difference due to the experiments. The first method to be discussed will be that of obtaining a probable error for each check and for each variety or treatment, and the data in Table 125, page 453, will be used to illustrate this method. The data are the yields obtained from a variety or strain test in which there were 10 plots of each strain.

The first step is to obtain the probable error for each set of ten checks, using either Bessel's or Peters' formula. In this illustration

---

*Since the preparation of this manuscript additional suggestions for further refinement of the method have been presented by Fisher in the fifth edition of his book. Wishart is considering further suggestions to the method, and Snedecor has made application of these newer methods in a paper as yet unpublished.

TABLE 125

YIELDS OBTAINED FROM A VARIETY TEST OF OATS, IN WHICH THERE WERE TEN PLOTS OF EACH VARIETY. THE RESULTS ARE COMPARED WITH THE CALCULATED CHECK YIELDS

Row No.	Replicates										Average Bushels Per Acre	Calculated Check by Graded Method	Gain or Loss
0 Check	41.0	35.8	36.0	46.5	37.3	33.5	34.5	33.5	33.2	27.0	35.8 ± 1.10		
1	34.4	42.7	34.0	40.7	35.8	24.9	37.2	41.1	36.7	37.3	36.5 ± 1.06	35.5 ± 1.56	1.0 ± 1.89
2	34.5	34.2	43.0	40.0	39.5	23.5	36.6	34.2	43.9	39.5	36.9 ± 1.25	35.2 ± 1.55	1.7 ± 1.99
3	56.6	49.3	49.0	54.0	51.9	53.8	52.1	49.7	51.8	50.7	51.9 ± .51	34.9 ± 1.53	17.0 ± 1.61
4	42.8	36.9	45.2	42.2	40.2	38.2	38.5	43.0	47.1	38.0	41.2 ± .72	34.6 ± 1.52	6.6 ± 1.68
5 Check	31.5	31.8	40.2	38.0	44.0	27.4	34.4	31.2	24.1	40.3	34.3 ± 1.34		
6	40.5	40.4	32.7	35.6	42.0	40.0	35.4	37.8	42.4	37.2	38.4 ± .68	34.4 ± 1.51	4.0 ± 1.66
7	37.4	36.4	34.6	39.2	37.8	34.8	37.0	30.8	35.2	37.0	36.0 ± .50	34.4 ± 1.51	1.6 ± 1.59
8	38.9	33.2	38.4	44.1	34.6	36.9	39.2	40.2	32.0	34.5	37.2 ± .79	34.5 ± 1.51	2.7 ± 1.70
9	42.2	42.0	39.1	36.6	34.2	42.5	37.1	39.8	47.0	33.3	39.4 ± .90	34.5 ± 1.51	4.9 ± 1.76
10 Check	29.5	23.0	38.3	48.4	36.5	34.0	30.5	29.4	26.3	49.8	34.6 ± 1.90		
11	48.9	44.9	36.0	36.2	40.9	43.2	39.0	41.8	41.3	37.7	41.0 ± .86	35.1 ± 1.54	5.9 ± 1.76
12	41.7	38.4	40.0	39.0	36.6	30.1	33.4	37.9	40.0	36.2	37.3 ± .74	35.6 ± 1.56	1.7 ± 1.73
13	42.0	36.1	36.6	36.7	41.5	38.0	37.0	40.0	38.7	38.4	38.5 ± .44	36.0 ± 1.58	2.5 ± 1.64
14	33.9	36.3	34.8	35.5	34.5	46.7	41.2	38.4	45.3	36.3	38.3 ± .98	36.5 ± 1.60	1.8 ± 1.88
15 Check	38.1	21.5	47.7	46.6	40.8	29.5	32.1	34.1	28.7	51.4	37.0 ± 2.05		
16	42.5	47.3	37.8	36.3	38.6	36.6	34.5	37.0	46.7	32.2	38.9 ± 1.07	36.8 ± 1.62	2.1 ± 1.94
17	38.2	35.9	38.4	37.0	41.0	39.6	37.4	39.8	34.2	39.5	38.1 ± .43	36.5 ± 1.60	1.6 ± 1.66
18	31.1	38.3	34.8	41.2	34.0	38.1	35.1	33.0	38.9	35.4	36.0 ± .65	36.3 ± 1.59	− .3 ± 1.72
19	39.7	41.8	36.1	38.0	39.4	48.8	38.4	38.7	48.8	38.0	40.8 ± .95	36.0 ± 1.58	4.8 ± 1.84
20 Check	32.7	26.7	47.1	49.7	37.4	32.0	31.0	31.5	24.9	45.0	35.8 ± 1.85		
21	37.6	48.0	34.9	34.0	40.8	38.5	34.2	37.4	36.8	32.0	37.4 ± .96	36.3 ± 1.59	1.1 ± 1.86
22	26.4	43.4	43.0	37.3	38.0	37.5	36.7	32.6	34.4	33.2	36.2 ± 1.07	36.8 ± 1.62	− .6 ± 1.94
23	34.8	38.1	34.8	42.6	39.3	41.3	36.8	40.8	39.0	38.3	38.6 ± .56	37.3 ± 1.64	1.3 ± 1.73
24	42.8	47.4	42.5	34.6	34.0	38.1	42.5	37.5	40.0	36.0	39.5 ± .91	37.8 ± 1.66	1.7 ± 1.89
25 Check	33.4	31.7	45.0	48.5	47.5	29.7	32.6	33.5	30.5	51.0	38.3 ± 1.82		
26	33.7	44.3	37.6	31.3	33.4	49.5	42.4	35.2	39.0	32.1	37.8 ± 1.27	39.6 ± 1.74	− 1.8 ± 2.15
27	41.0	46.6	40.6	38.4	37.5	41.4	40.2	41.6	40.8	36.7	40.5 ± .58	40.9 ± 1.80	− .4 ± 1.89
28	38.4	22.9	36.4	28.5	22.4	29.7	31.6	34.9	27.0	34.3	30.6 ± 1.17	42.1 ± 1.85	−11.5 ± 2.19
29	33.8	21.3	29.2	34.0	33.3	34.6	48.0	33.3	42.4	29.6	33.9 ± 1.54	43.4 ± 1.91	− 9.5 ± 2.45
30 Check	49.5	42.3	41.0	37.4	45.7	48.5	57.3	44.5	46.2	34.7	44.7 ± 1.37		

Probable error of checks in per cent = 4.39

Bessel's formula has been used. In experiments of this kind there are usually a large number of strains under test and a large number of check plots. By summing the average yields and the probable errors for all the groups of checks in the experiment and dividing the total sum of the errors by the total sum of the average yields of the check plots we obtain a probable error in per cent. This is considered to be the probable error for the check plots and is the probable error that is used for the calculated check yields. The calculated, or theoretical, check is obtained by the graded method and this calculated check is multiplied by the probable error in per cent. For the data in Table 125 the average probable error of the mean of ten check plots is found to be 4.39 per cent. The calculated check yield for row 1 is 35.5 bushels, and multiplying 35.5 by 4.39 we have a probable error of 1.56 bushels. The probable errors for the remaining calculated check yields are obtained in the same way.

The yields of the different varieties with their probable errors are compared with the theoretical check yields and the gains or losses obtained, as shown in the table. The probable error of the gain or loss is determined by the usual method for the probable error of a difference. For example, for the variety in row 1 the average yield is 36.5 ± 1.06 and the calculated check yield is 35.5 ± 1.56. The difference with its probable error is 1.0 ± 1.89. The other comparisons are made in a similar way.

After determining the difference, or gain or loss, and its probable error, we may interpret the results on the basis of whether the difference is significantly larger than its probable error. As explained before, unless the difference is more than three (about 3.2) times the probable error of the difference it is usually unwise to conclude that the result is significant.

In the case just illustrated the probable error of the difference has been obtained without considering the correlation. For more exact determinations the effect of correlation should be eliminated, but since this method is used primarily as a guide to indicate what varieties are worthy of being continued in the test it saves labor to make the calculations as indicated rather than to eliminate the effect of correlation. If the correlation is to be considered, it is necessary to determine the calculated check yield for each plot rather than for the mean of the ten plots, so as to obtain the differences pair by pair. This increases the amount of work considerably, and if one keeps in mind the fact that in such experiments if the correlation were considered the probable error of the difference would usually be lower, therefore certain comparisons that border on significance may be significant when the effect of correlation is considered.

In order to show that under the usual conditions this method of interpreting results, especially the use of the graded method, will not lead to a serious error, the following data will be of interest.

Table 126

Results Obtained by Correlating the Yield and Gain or Loss as Compared with the Calculated Check. The Data for Each Year Represent the Results for All Varieties Tested

Wheat	
1927	.919 ± .010
1928	.960 ± .007
1929	.814 ± .021
1931	.755 ± .027

Oats	
1927	.749 ± .026
1928	.767 ± .028
1929	.796 ± .029
1931	.592 ± .035

The results given in Table 126 show the correlation between the yields and the gains or losses as compared with the calculated checks, for all varieties in the test. The correlations are all high and significant, showing that there is a very close relation between yield and gain or loss as compared with the check. Further information is furnished by the data in Table 127.

TABLE 127

COEFFICIENTS OF CORRELATION OBTAINED BETWEEN THE SAME VARIETIES GROWN IN DIFFERENT YEARS

CROP	CORRELATION BETWEEN	COEFFICIENT OF CORRELATION	
		YIELD IN BUSHELS PER ACRE	GAIN OR LOSS IN COMPARISON WITH CHECK
Oats			
	1925 and 1926	-.123 ± .077	.223 ± .074
	1926 and 1927	.299 ± .083	.610 ± .057
	1927 and 1928	.354 ± .072	.432 ± .067
	1928 and 1929	.595 ± .052	.514 ± .059
	1929 and 1930	.204 ± .088	.229 ± .087
	1930 and 1931	.422 ± .076	.359 ± .081
	1931 and 1932	.073 ± .068	.231 ± .065
Wheat			
	1925 and 1927	.333 ± .097	.503 ± .082
Barley			
	1924 and 1925	.603 ± .075	.422 ± .097
	1926 and 1927	.373 ± .066	.514 ± .056
	1927 and 1928	.392 ± .071	.495 ± .063

The results in Table 127 show the correlation between the actual yields for the same varieties for different years and the correlation between the same varieties by using the gain or loss in comparison with the check. It is seen that the average correlation for the gain or loss is a little higher than the correlation obtained from the actual yields. If any error were introduced we would not obtain the higher correlation so often, but due to errors of sampling we might expect very low or even negative correlation in some instances. This indicates that the method of graded checks does not necessarily introduce an error.

The method for calculating probable errors just described is somewhat laborious, and simpler methods may be used. A second method is one in which a general probable error for the experiment is determined by calculating the probable errors for the check plots only. For example, for the data in Table 125 the average probable error for the check plots is 4.39 per cent. This probable error may be used as a general probable error for comparing varieties with each other or with their calculated check yields.

Since we are concerned with differences we obtain the probable error of the difference from $4.39\sqrt{2}$, giving 6.21. For selecting the better yielding varieties we may use two or three times this probable error of the difference, depending on the standard for selection which it is desired to follow, and multiply the mean yields of the varieties by this value. Adopting three times the probable error of the difference as the standard, for the data at hand we have 6.21×3, or 18.63. The mean yields of the varieties are multiplied by this value and if the gain as compared with the check is greater than this amount, we would conclude that the variety is significantly better than the check variety. For example, the variety in row 3 of Table 125 has an average yield of 51.9 bushels. Multiplying 51.9 by three times the probable error of the difference in per cent, or 18.63, we obtain 9.67. The gain of this variety over the check is 17.0, which is greater than 9.67, and we therefore conclude that the variety in row 3 is significantly better than the check variety. Other comparisons may also be made.

This method requires less time for making the calculations and it is satisfactory as a simple method for determining the superior or inferior varieties in a test. One objection to this method is that the probable error is based on the variability of one variety only and this variety may not be representative of all the varieties.

Another method, suggested by Hayes, is the deviation from the mean method, and by this method also a general probable error for the experiment is determined. For all the varieties in the test the deviation between each plot of a variety and its own mean is obtained and these deviations are squared and summed. The method may be illustrated with the data in Table 125, using the varieties only.

For the variety in row 1 the first plot yields 34.4 bushels and the mean for this variety is 36.5. The deviation is therefore 34.4–36.5, or –2.1. For the second plot of this variety the deviation is 42.7–36.5, or 6.2, and the other deviations for this variety are obtained in a similar manner. The deviations for the other varieties are obtained by taking the deviation between each plot of the variety and its particular mean. The steps are as follows:

	YIELD—MEAN	D	D^2
For variety in row 1	34.4–36.5 42.7–36.5 34.0–36.5 etc.	–2.1 6.2 –2.5	4.41 38.44 6.25
For variety in row 2	34.5–36.9 34.2–36.9 43.0–36.9 etc.	–2.4 –2.7 6.1	5.76 7.29 37.21
For variety in row 3	56.6–51.9 49.3–51.9 49.0–51.9 etc.	4.7 2.6 2.9	22.09 6.76 8.41

Squaring and summing all of the deviations for the 24 varieties in the test we have 3884.50.

The probable error of a single plot is obtained from the formula

$$P.E._s = \pm .6745\sqrt{\frac{\Sigma D^2 n}{N\,(n-1)}}$$

in which N is the total number of plots and n is the number of plots of each variety. For the 24 varieties in Table 125 there are 10 plots of each variety and therefore a total of 240 plots, and $n=10$ and $N=240$. Substituting the necessary values in the formula we have

$$P.E._s = \pm .6745\sqrt{\frac{3884.50\times 10}{240\times 9}} = \pm 2.861$$

This probable error should be expressed in per cent, which is done by dividing the probable error by the mean yield of all the varieties and multiplying by 100. We have

$$\frac{2.861}{38.371}\times 100 = 7.46 \text{ per cent}$$

The probable error of the mean of ten plots is $7.46/\sqrt{10}$, or 2.36 per cent. This is the general probable error which will be used in the interpretation of the results, and it may be applied as described for the probable error obtained from the checks. Thus the probable error of the difference between the means of two varieties is $2.36\sqrt{2}$, and this may be multiplied by two or three, depending on the standard adopted for selection. Varieties may be compared with each other or with the calculated checks, as with the preceding method.

In most cases there will probably not be a great deal of difference between the general probable error obtained from the check plots alone and the probable error obtained by the deviation from the mean method. Occasionally there may be considerable variation but, as Hayes shows, the average probable error of a single determination, or plot, is about the same by either method. Some of the comparisons are shown by the following data, presented by Hayes.

CROP	NUMBER OF YEARS AVERAGED	*P. E.*, FROM CHECKS	*P. E.*, DEVIATION FROM MEAN
Spring wheat	6	9.3	9.9
Oats	6	7.7	6.9
Barley	6	9.4	9.2

The results obtained are very similar and either method may be used in selecting strains to be continued in a test. The method of determining the probable error from the check plots only is less laborious than the deviation from the mean method.

Another method for determining a general probable error for an experiment is suggested by 'Student' and is in effect the application of the method of variance. This method may be illustrated by the data in Table 128.

TABLE 128

METHOD FOR DETERMINING A GENERALIZED PROBABLE ERROR FOR AN EXPERIMENT AFTER ELIMINATING THE EFFECTS OF VARIETIES OR TREATMENT AND OF REPLICATES

VARIETIES	REPLICATES					TOTAL	MEAN
	1	2	3	4	5		
A	40	46	44	38	42	210	42
B	38	44	42	36	40	200	40
C	36	40	38	34	42	190	38
D	34	38	36	32	30	170	34
E	42	42	40	40	46	210	42
TOTAL	190	210	200	180	200	980	GENERAL MEAN
MEAN	38	42	40	36	40		39.2

TOTAL DEVIATION OBTAINED BY FINDING THE DEVIATION BETWEEN EACH PLOT AND THE MEAN OF ALL PLOTS, SQUARING, AND SUMMING	
D	*D*²
.8	.64
6.8	46.24
4.8	23.04
-1.2	1.44
2.8	7.84
-1.2	1.44
4.8	23.04
2.8	7.84
-3.2	10.24
.8	.64
-3.2	10.24
.8	.64
-1.2	1.44
-5.2	27.04
2.8	7.84
-5.2	27.04
-1.2	1.44
-3.2	10.24
-7.2	51.84
-9.2	84.64
2.8	7.84
2.8	7.84
.8	.64
.8	.64
6.8	46.24
	$\Sigma=408.00$ $\sigma_T^2=16.32$

SUM OF SQUARES FOR VARIETIES		SUM OF SQUARES FOR REPLICATES	
D	*D*²	*D*	*D*²
2.8	7.84	-1.2	1.44
.8	.64	2.8	7.84
-1.2	1.44	.8	.64
-5.2	27.04	-3.2	10.24
2.8	7.84	.8	.64
	$\Sigma=44.80$ $\sigma_V^2=8.96$		$\Sigma=20.80$ $\sigma_R^2=4.16$

$$P.E._s=\pm.6745\sqrt{\frac{5\times5(16.32-8.96-4.16)}{4\times4}}=\pm1.5$$

The formula is

$$P.E._s = \pm .6745 \sqrt{\frac{Nn\,(\sigma_T^2 - \sigma_V^2 - \sigma_R^2)}{(N-1)\,(n-1)}}$$

in which

σ_T^2 = variance for the yields of all the individual plots in the test.

σ_V^2 = variance of the means of the varieties

σ_R^2 = variance of the means of the replicates

N = number of varieties

n = number of plots of each variety.

These constants are obtained by calculating as follows:

$\sigma_T^2 = \frac{\Sigma P^2}{Nn} - M^2$ where P = yields of individual plots

M = general mean of all plots

$\sigma_V^2 = \frac{\Sigma V^2}{N} - M^2$ where V = mean yields of varieties

$\sigma_R^2 = \frac{\Sigma R^2}{n} - M^2$ where R = mean yields of replicates

These formulas are to be used when working from an assumed mean of zero, or when the individual items themselves are squared. When convenient the variances may be determined by obtaining the differences between the individual items and the general mean, M. The latter method has been followed for the values in Table 128. The method to be followed is merely a matter of convenience or ease in calculation.

Obtaining the necessary values and substituting in the formula, we have

$$P.E._s = \pm .6745 \sqrt{\frac{5 \times 5\,(16.32 - 8.96 - 4.16)}{4 \times 4}} = \pm 1.5$$

Expressing this probable error in per cent we have

$$\frac{1.5}{39.2} \times 100, \text{ or } 3.8 \text{ per cent}$$

This is the probable error in per cent for a single plot. The probable error of the mean of five plots is $3.8/\sqrt{5}$, or 1.7 per cent.

The deviation from the mean method for these same data gives a probable error for a single plot of 5.1 per cent, and the probable error of the mean of five plots is 2.3 per cent. There is a reduction in the probable error by eliminating the effect of varieties and replicates from the total variation. This method is satisfactory, but if there are a large number of varieties under test considerable calculation is necessary. For experiments like soil fertility studies and cultural and rotation experiments the method may be used, but the plant breeder is often concerned with comparing the yielding ability of several hundred different strains or varieties, and in such cases it is better for the preliminary analysis to use some method that may be quickly applied.

The foregoing methods are suggested as those that may be applied for the analysis of the results for any one particular year. For the analysis of results for several years a different technic may be used. For example, if a system of testing with frequent check plots has been followed and the comparison between the test plots and the check plots has been made, since each result has been compared with the same check it is possible to compare these gain or loss values directly with each other by means of a method such as 'Student's' or Fisher's, as described in Chapter XII. Suppose that in a variety test conducted for a period of five years two varieties when compared with the check showed the following gains.

	GAIN OVER Check	
YEAR	VARIETY A	VARIETY B
1928	7.6	4.2
1929	4.3	4.0
1930	5.8	3.0
1931	8.7	4.2
1932	4.0	2.1

These results may be compared by 'Student's' method to show whether there is a significant gain over the check for either variety. If it is desired to compare the two varieties directly, then the differences between the two series may be obtained pair by pair and the results evaluated by 'Student's' or Fisher's method. As explained in Chapter XII, if one should have say the results for five years for one comparison and perhaps for only four years for another comparison, then Fisher's method applicable to such cases should be used.

The method of variance analysis as discussed fully in chapters XIII and XIV may be used for the evaluation of results, especially for soil fertility and cultural experiments. In such cases it will be possible to combine all the tests and make one general analysis, along lines similar to those given in detail for the analysis of a complex experiment.

We may conclude this discussion by stating that for the results for a single year some method that will enable one to judge the performance of the different varieties should be applied, and any one of those described above may be used. For the final analysis the results for the different years should be brought together and evaluated by Fisher's or 'Student's' method for determining the significance of results, or by the application of the analysis of variance.

Recommendations Regarding Plot Arrangement. There are two general types of experiment for which field plots are used. One is the experiment conducted by the plant breeder when comparing new strains obtained either from selection or by hybridization, and the other is in connection with soil fertility studies and cultural and rotation experiments. Experience has shown that for the first type of experiment small plots replicated many times are more satisfactory than larger plots with fewer replications.

For the small grains, as wheat, barley, and rice, small plots one row in width and from 12 to 18 feet in length have been found to be very satisfactory, especially for the test in the first few years. For the final comparison 3-row blocks from 12 to 18 feet in length are very satisfactory. If competition has a serious effect, then the single-row plots may not be so satisfactory for even the earlier tests, and it may be better to use 3-row blocks. As shown earlier, competition may be present but may not have any serious effect so far as yield is concerned, and it seems to be more satisfactory to use more replications of the single-row plots than to have fewer replications of the 3-row blocks, especially in the preliminary tests.

For other crops, such as corn, potatoes, kaoliang, and cotton, some modification of the single-row plot may be used. The rows should be longer and the distance between the rows must be greater, depending on the crop being studied. A row long enough to include from 20 to 30 plants is satisfactory provided the rows can be replicated several times. Four replications for the more careful tests is a minimum, and it is better to have as many as nine replications, if possible.

Regarding the number of replications, it is well to have more replications than may have been indicated from a study of the blank test. For example, if the analysis from the blank test shows that six replications are sufficient, it is better to have more than six in order to provide for loss through accident. If competition seems to be serious, blocks of three or more rows are recommended for all testing work. Since this necessitates a larger amount of land and if land is a limiting factor, it will be found that strains similar in their general characteristics may be grown in single-row plots and competition may be avoided by placing the rows farther apart. This may be objected to since it may be thought that the greater distance between rows will mean that the plants are grown under different conditions than would normally be the case. This is not a particularly serious objection since the important point is to determine the best strain, and by this method it is possible to eliminate the poorer varieties. The final elimination may be made following a more careful test, as in 3-row blocks, avoiding any effect of competition by considering the middle rows only.

For such experiments as soil fertility studies and cultural and rotation experiments, it is desirable to have small plots replicated several times rather than to have larger plots with only one or two replications. While small plots are recommended, it is important that the plots be large enough to include a sufficient number of plants. This is not a problem with crops like rice, wheat, and the like, but it is important with the larger crops, such as cotton or kaoliang. The plots may be arranged at random and there should be at least five plots of each treatment. There should be at least one control plot in each replication. If the plots are arranged in systematic order, then a check or control plot should be placed every third, fourth, or fifth plot, depending on the amount of land available for the experiment.

The manner of handling the plots needs also to be considered when deciding on the size and shape of plots. If the plots are large so that they require preparation by machinery, then it is important that the plots be arranged with a sufficient border between them so that soil will not be moved or dragged from one plot to another. It has already been suggested that border strips should be left between the plots so as to offset any effect of competition due to different amounts of fertilizer, and the like. For soil fertility studies with rice, a plot 10 by 20 feet, giving an area 1/30 of a mow, may be used. This will permit of the preparation of the plots by hand and thus prevent any effect of movement of soil from one plot to another. With any crop which is to be irrigated, it is desirable to have well-built dikes so that there will be no water moving from one plot to another. These should be made of the same soil as that around the plots and there should be a border strip between the plots planted to the same crop.

For plots that are to remain under test for many years it is important to have a preliminary test on the plots for two or more crops, and, as explained earlier in this chapter, the results from these preliminary tests may be of value in the interpretation of the results from the experiment. The preliminary cropping is also of value to indicate whether all of the plots in the field chosen are suitable for experimental work. Following the preliminary cropping the location of the different treatments may be arranged at random, or in some suitable way that will enable the experimenter to make accurate determinations from the treatments in question.

The number of replications needed for experiments in soil fertility will depend in part on the differences to be measured, but at least four to nine replications of small plots, such as 1/30 or 1/50 of a mow, will be desirable. If the plots are large it is desirable to have at least three or four replications.

It is impossible to make definite rules which must be followed under all conditions, as the kind of plot, number of replications, and the like will depend somewhat on land, labor, and other facilities available. One should determine the best arrangement for the conditions under which he has to work, and then the experiments should be continued for a sufficient length of time to make certain that the results obtained are reliable. In order to meet certain needs, the following specific suggestions are made as to methods suitable for variety testing. For such crops as wheat, rice, barley, millet, and the like, the plan may be as follows:

YEAR	KIND OF PLOT	LENGTH OF ROW	NUMBER OF REPLICATIONS	FREQUENCY OF CHECK
First year	Head or plant rows	3 to 6 feet	0	Every tenth
Second year	Single rows	12 to 18 feet	1 or 2	Every fifth
Third year	Single rows	12 to 18 feet	4 or more	Every fifth
Fourth year	Single rows	12 to 18 feet	9 or more	Every fifth
Fifth year	3-row blocks	12 to 18 feet	9 or more	Every third
	Drilled plots of 5 to 7 rows	30 to 60 feet	5 or more	Every third

The distance between the rows may be from 1 to 2 feet, depending on the crop used. If sufficient land and labor are available, 3-row blocks may be used earlier in the test.

For such crops as beans and sesamum, the plan outlined above may be followed, but the distance between rows will be from 2 or 2 1/2 to 3 feet. For potatoes (sweet or Irish), kaoliang, corn, and cotton, the above plan may be followed in general, but the rows will be from 20 to 50 or more feet in length and from 2 to 3 feet in width. In order to avoid competition, it may be preferred to use 3-row blocks earlier in the test, especially if widely differing strains are being tested. General suggestions relative to the size and to the number of replications of plots to be used for rotation and soil experiments, and the like, have been made earlier in this chapter.

APPENDIX

TABLE I

SUMS OF POWERS OF NATURAL NUMBERS

n	$S(n)$	$S(n^2)$	$S(n^3)$	$S(n^4)$	n
1	1	1	1	1	1
2	3	5	9	17	2
3	6	14	36	98	3
4	10	30	100	354	4
5	15	55	225	979	5
6	21	91	441	2275	6
7	28	140	784	4676	7
8	36	204	1296	8772	8
9	45	285	2025	15333	9
10	55	385	3025	25333	10
11	66	506	4356	39974	11
12	78	650	6084	60710	12
13	91	819	8281	89271	13
14	105	1015	11025	127687	14
15	120	1240	14400	178312	15
16	136	1496	18496	243848	16
17	153	1785	23409	327369	17
18	171	2109	29241	432345	18
19	190	2470	36100	562666	19
20	210	2870	44100	722666	20
21	231	3311	53361	917147	21
22	253	3795	64009	1151403	22
23	276	4324	76176	1431244	23
24	300	4900	90000	1763020	24
25	325	5525	105625	2153645	25

TABLE II

TABLED VALUES TO FACILITATE THE FITTING OF A LOGARITHMIC CURVE OF THE GENERAL FORMULA $y = a + bx + c \log x$

x	$\log x$	$S(\log x)$	$x \log x$	$S(x \log x)$	$(\log x)^2$	$S(\log x)^2$
1	.0000	.0000	.0000	.0000	.0000	.0000
2	.3010	.3010	.6020	.6020	.0906	.0906
3	.4771	.7781	1.4313	2.0333	.2276	.3182
4	.6021	1.3802	2.4084	4.4417	.3625	.6807
5	.6990	2.0792	3.4950	7.9367	.4886	1.1693
6	.7782	2.8574	4.6692	12.6059	.6056	1.7749
7	.8451	3.7025	5.9157	18.5216	.7142	2.4891
8	.9031	4.6056	7.2248	25.7464	.8156	3.3047
9	.9542	5.5598	8.5878	34.3342	.9105	4.2152
10	1.0000	6.5598	10.0000	44.3342	1.0000	5.2152
11	1.0414	7.6012	11.4554	55.7896	1.0845	6.2997
12	1.0792	8.6804	12.9504	68.7400	1.1647	7.4644
13	1.1139	9.7943	14.4807	83.2207	1.2408	8.7052
14	1.1461	10.9404	16.0454	99.2661	1.3135	10.0187
15	1.1761	12.1165	17.6415	116.9076	1.3832	11.4019
16	1.2041	13.3206	19.2656	136.1732	1.4499	12.8518
17	1.2304	14.5510	20.9168	157.0900	1.5139	14.3657
18	1.2553	15.8063	22.5954	179.6854	1.5758	15.9415
19	1.2788	17.0851	24.2972	203.9826	1.6353	17.5768
20	1.3010	18.3861	26.0200	230.0026	1.6926	19.2694
21	1.3222	19.7083	27.7662	257.7688	1.7482	21.0176
22	1.3424	21.0507	29.5328	287.3016	1.8020	22.8196
23	1.3617	22.4124	31.3191	318.6207	1.8542	24.6738
24	1.3802	23.7926	33.1248	351.7455	1.9050	26.5788
25	1.3979	25.1905	34.9475	386.6930	1.9541	28.5329
26	1.4150	26.6055	36.7900	423.4830	2.0022	30.5351
27	1.4314	28.0369	38.6478	462.1308	2.0489	32.5840
28	1.4472	29.4841	40.5216	502.6524	2.0944	34.6784
29	1.4624	30.9465	42.4096	545.0620	2.1386	36.8170
30	1.4771	32.4236	44.3130	589.3750	2.1818	38.9988
31	1.4914	33.9150	46.2334	635.6084	2.2243	41.2231
32	1.5051	35.4201	48.1632	683.7716	2.2653	43.4884
33	1.5185	36.9386	50.1105	733.8821	2.3058	45.7942
34	1.5315	38.4701	52.0710	785.9531	2.3455	48.1397
35	1.5441	40.0142	54.0435	839.9966	2.3842	50.5239
36	1.5563	41.5705	56.0268	896.0234	2.4221	52.9460
37	1.5682	43.1387	58.0234	954.0468	2.4593	55.4053
38	1.5798	44.7185	60.0324	1,014.0792	2.4958	57.9011
39	1.5911	46.3096	62.0529	1,076.1321	2.5316	60.4327
40	1.6021	47.9117	64.0840	1,140.2161	2.5667	62.9994

TABLE II—*Continued*

x	*log x*	*S (log x)*	*x log x*	*S (x log x)*	$(log\ x)^2$	$S\ (log\ x)^2$
41	1.6128	49.5245	66.1248	1,206.3409	2.6011	65.6005
42	1.6232	51.1477	68.1744	1,274.5153	2.6348	68.2355
43	1.6335	52.7812	70.2405	1,344.7558	2.6683	70.9036
44	1.6435	54.4247	72.3140	1,417.0698	2.7011	73.6047
45	1.6532	56.0779	74.3940	1,491.4638	2.7331	76.3378
46	1.6628	57.7407	76.4888	1,567.9526	2.7649	79.1027
47	1.6721	59.4128	78.5887	1,646.5413	2.7959	81.8986
48	1.6812	61.0940	80.6976	1,727.2389	2.8264	84.7250
49	1.6902	62.7842	82.8198	1,810.0587	2.8568	87.5818
50	1.6990	64.4832	84.9500	1,895.0087	2.8866	90.4684
51	1.7076	66.1908	87.0876	1,982.0962	2.9159	93.3843
52	1.7160	67.9068	89.2320	2,071.3283	2.9447	96.3290
53	1.7243	69.6311	91.3879	2,162.7162	2.9732	99.3022
54	1.7324	71.3635	93.5496	2,256.2658	3.0012	102.3034
55	1.7404	73.1039	95.7220	2,351.9878	3.0290	105.3324
56	1.7482	74.8521	97.8992	2,449.8870	3.0562	108.3886
57	1.7559	76.6080	100.0863	2,549.9733	3.0832	111.4718
58	1.7634	78.3714	102.2772	2,652.2505	3.1096	114.5814
59	1.7709	80.1423	104.4831	2,756.7336	3.1361	117.7175
60	1.7782	81.9205	106.6920	2,863.4256	3.1620	120.8795
61	1.7853	83.7058	108.9033	2,972.3289	3.1873	124.0668
62	1.7924	85.4982	111.1288	3,083.4577	3.2127	127.2795
63	1.7993	87.2975	113.3559	3,196.8136	3.2375	130.5170
64	1.8062	89.1037	115.5968	3,312.4104	3.2624	133.7794
65	1.8129	90.9166	117.8385	3,430.2489	3.2866	137.0660
66	1.8195	92.7361	120.0870	3,550.3359	3.3106	140.3766
67	1.8261	94.5622	122.3487	3,672.6846	3.3346	143.7112
68	1.8325	96.3947	124.6100	3,797.2946	3.3581	147.0693
69	1.8388	98.2335	126.8772	3,924.1718	3.3812	150.4505
70	1.8451	100.0786	129.1570	4,053.3288	3.4044	153.8549
71	1.8513	101.9299	131.4423	4,184.7711	3.4273	157.2822
72	1.8573	103.7872	133.7256	4,318.4967	3.4496	160.7318
73	1.8633	105.6505	136.0209	4,454.5176	3.4719	164.2037
74	1.8692	107.5197	138.3208	4,592.8384	3.4939	167.6976
75	1.8751	109.3948	140.6325	4,733.4709	3.5160	171.2136
76	1.8808	111.2756	142.9408	4,876.4117	3.5374	174.7510
77	1.8865	113.1621	145.2605	5,021.6722	3.5589	178.3099
78	1.8921	115.0542	147.5838	5,169.2560	3.5800	181.8899
79	1.8976	116.9518	149.9104	5,319.1664	3.6009	185.4908
80	1.9031	118.8549	152.2480	5,471.4144	3.6218	189.1126
81	1.9085	120.7634	154.5885	5,626.0029	3.6424	192.7550
82	1.9138	122.6772	156.9316	5,782.9345	3.6626	196.4176
83	1.9191	124.5963	159.2853	5,942.2198	3.6829	200.1005
84	1.9243	126.5206	161.6412	6,103.8610	3.7029	203.8034
85	1.9294	128.4500	163.9990	6,267.8600	3.7226	207.5260

TABLE II—*Continued*

x	$\log x$	$S(\log x)$	$x \log x$	$S(x \log x)$	$(\log x)^2$	$S(\log x)^2$
86	1.9345	130.3845	166.3670	6,434.2270	3.7423	211.2683
87	1.9395	132.3240	168.7365	6,602.9635	3.7617	215.0300
88	1.9445	134.2685	171.1160	6,774.0795	3.7811	218.8111
89	1.9494	136.2179	173.4966	6,947.5761	3.8002	222.6113
90	1.9542	138.1721	175.8780	7,123.4541	3.8189	226.4302
91	1.9590	140.1311	178.2690	7,301.7231	3.8377	230.2679
92	1.9638	142.0949	180.6696	7,482.3927	3.8565	234.1244
93	1.9685	144.0634	183.0705	7,665.4632	3.8750	237.9994
94	1.9731	146.0365	185.4714	7,850.9346	3.8931	241.8925
95	1.9777	148.0142	187.8815	8,038.8161	3.9113	245.8038
96	1.9823	149.9965	190.3008	8,229.1169	3.9295	249.7333
97	1.9868	151.9833	192.7196	8,421.8365	3.9474	253.6807
98	1.9912	153.9745	195.1376	8,616.9741	3.9649	257.6456
99	1.9956	155.9701	197.5644	8,814.5385	3.9824	261.6280
100	2.0000	157.9701	200.0000	9,014.5385	4.0000	265.6280

TABLE III

TABLE GIVING CORRESPONDING VALUES FOR r COMPUTED FROM r_r

r_r	r	r_r	r	r_r	r
.10	.105	.40	.416	.70	.717
.11	.115	.41	.426	.71	.726
.12	.126	.42	.436	.72	.736
.13	.136	.43	.446	.73	.746
.14	.146	.44	.457	.74	.756
.15	.157	.45	.467	.75	.765
.16	.167	.46	.477	.76	.775
.17	.178	.47	.487	.77	.785
.18	.188	.48	.497	.78	.794
.19	.199	.49	.508	.79	.804
.20	.209	.50	.518	.80	.813
.21	.219	.51	.528	.81	.823
.22	.230	.52	.538	.82	.833
.23	.240	.53	.548	.83	.842
.24	251	.54	.558	.84	.852
.25	261	.55	.568	.85	.861
.26	.271	.56	.578	.86	.870
.27	.282	.57	.588	.87	.880
.28	.292	.58	.598	.88	.889
.29	.303	.59	.608	.89	.899
.30	.313	.60	.618	.90	.908
.31	.323	.61	.628	.91	.917
.32	.333	.62	.638	.92	.927
.33	.344	.63	.648	.93	.936
.34	.354	.64	.658	.94	.945
.35	.364	.65	.668	.95	.954
.36	375	.66	.677	.96	.964
.37	.385	.67	.687	.97	.973
.38	.395	.68	.697	.98	.982
.39	.406	.69	.707	.99	.991

TABLE IV

VALUES FOR FACILITATING COMPUTATION OF THE PROBABLE ERROR OF A SINGLE OBSERVATION, FROM BESSEL'S FORMULA

VALUE OF THE FACTOR $\frac{.6745}{\sqrt{N-1}}$ FOR N 1 TO 99

N		0	1	2	3	4	5	6	7	8	9
1		.2248	.2133	.2034	.1947	.1871	.1803	.1742	.1686	.1636	.1590
2	.6745	.1547	.1508	.1472	.1438	.1406	.1377	.1349	.1323	.1298	.1275
3	.4769	.1253	.1231	.1211	.1192	.1174	.1157	.1140	.1124	.1109	.1094
4	.3894	.1080	.1066	.1053	.1041	.1029	.1017	.1005	.0995	.0984	.0974
5	.3372	.0964	.0954	.0944	.0935	.0926	.0918	.0909	.0901	.0893	.0886
6	.3016	.0878	.0871	.0864	.0857	.0850	.0843	.0837	.0830	.0824	.0818
7	.2754	.0812	.0806	.0800	.0795	.0789	.0784	.0779	.0774	.0769	.0764
8	.2549	.0759	.0754	.0749	.0745	.0740	.0736	.0732	.0727	.0723	.0719
9	.2385	.0715	.0711	.0707	.0703	.0699	.0696	.0692	.0688	.0685	.0681

VALUES FOR FACILITATING COMPUTATION OF THE PROBABLE ERROR OF THE MEAN, FROM BESSEL'S FORMULA

VALUE OF THE FACTOR $\frac{.6745}{\sqrt{N(N-1)}}$ FOR N 1 TO 99

N		0	1	2	3	4	5	6	7	8	9
1		.0711	.0643	.0587	.0540	.0500	.0465	.0435	.0409	.0386	.0365
2	.4769	.0346	.0329	.0314	.0300	.0287	.0275	.0265	.0255	.0245	.0237
3	.2754	.0229	.0221	.0214	.0208	.0201	.0196	.0190	.0185	.0180	.0175
4	.1947	.0171	.0167	.0163	.0159	.0155	.0152	.0148	.0145	.0142	.0139
5	.1508	.0136	.0134	.0131	.0128	.0126	.0124	.0122	.0119	.0117	.0115
6	.1231	.0113	.0111	.0110	.0108	.0106	.0105	.0103	.0101	.0100	.0098
7	.1041	.0097	.0096	.0094	.0093	.0092	.0091	.0089	.0088	.0087	.0086
8	.0901	.0085	.0084	.0083	.0082	.0081	.0080	.0079	.0078	.0077	.0076
9	.0795	.0075	.0075	.0074	.0073	.0072	.0071	.0071	.0070	.0069	.0068

TABLE V

VALUES FOR FACILITATING COMPUTATION OF THE PROBABLE ERROR OF A SINGLE OBSERVATION, FROM PETERS' FORMULA

VALUE OF THE FACTOR $\frac{.8453}{\sqrt{N(N-1)}}$ FOR N 1 TO 99

N		0	1	2	3	4	5	6	7	8	9
1		.0891	.0806	.0736	.0677	.0627	.0583	.0546	.0513	.0483	.0457
2	.5977	.0434	.0412	.0393	.0376	.0360	.0345	.0332	.0319	.0307	.0297
3	.3451	.0287	.0277	.0268	.0260	.0252	.0245	.0238	.0232	.0225	.0220
4	.2440	.0214	.0209	.0204	.0199	.0194	.0190	.0186	.0182	.0178	.0174
5	.1890	.0171	.0167	.0164	.0161	.0158	.0155	.0152	.0150	.0147	.0145
6	.1543	.0142	.0140	.0137	.0135	.0133	.0131	.0129	.0127	.0125	.0123
7	.1304	.0122	.0120	.0118	.0117	.0115	.0113	.0112	.0110	.0109	.0108
8	.1130	.0106	.0105	.0104	.0102	.0101	.0100	.0099	.0098	.0097	.0096
9	.0996	.0094	.0093	.0092	.0091	.0090	.0089	.0089	.0088	.0087	.0086

VALUES FOR FACILITATING COMPUTATION OF THE PROBABLE ERROR OF THE MEAN, FROM PETERS' FORMULA

VALUE OF THE FACTOR $\frac{.8453}{N\sqrt{N-1}}$ FOR N 1 TO 99

N		0	1	2	3	4	5	6	7	8	9
1		.0282	.0243	.0212	.0188	.0167	.0151	.0136	.0124	.0114	.0105
2	.4226	.0097	.0090	.0084	.0078	.0073	.0069	.0065	.0061	.0058	.0055
3	.1992	.0052	.0050	.0047	.0045	.0043	.0041	.0040	.0038	.0037	.0035
4	.1220	.0034	.0033	.0031	.0030	.0029	.0028	.0027	.0027	.0026	.0025
5	.0845	.0024	.0023	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0019
6	.0630	.0018	.0018	.0017	.0017	.0017	.0016	.0016	.0016	.0015	.0015
7	.0493	.0015	.0014	.0014	.0014	.0013	.0013	.0013	.0013	.0012	.0012
8	.0399	.0012	.0012	.0011	.0011	.0011	.0011	.0011	.0010	.0010	.0010
9	.0332	.0010	.0010	.0010	.0009	.0009	.0009	.0009	.0009	.0009	.0009

TABLE VI

TABLE FOR ESTIMATING PROBABILITY BASED ON THE NORMAL PROBABILITY INTEGRAL CORRESPONDING TO VALUES OF $\frac{x}{\sigma}$. TOTAL AREA OF THE CURVE IS ASSUMED TO BE 100000. x IS THE DISTANCE FROM THE MEAN AND σ IS THE STANDARD DEVIATION.*

$\frac{x}{\sigma}$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
0.00	00000	40	80	120	159	199	239	279	319	359
0.03	1197	1237	1276	1316	1356	1396	1436	1476	1516	1555
0.06	2392	2432	2472	2512	2551	2591	2631	2671	2711	2751
0.09	3586	3625	3665	3705	3744	3784	3824	3864	3903	3943
0.12	4776	4815	4855	4895	4934	4974	5013	5053	5093	5132
0.15	5962	6001	6041	6080	6119	6159	6198	6238	6277	6317
0.18	7142	7182	7221	7260	7299	7338	7378	7417	7456	7495
0.21	8317	8356	8395	8434	8473	8512	8551	8590	8628	8667
0.24	9483	9522	9561	9600	9638	9677	9716	9754	9793	9832
0.27	10642	10680	10719	10757	10796	10834	10872	10911	10949	10988
0.30	11791	11829	11867	11905	11943	11981	12019	12058	12096	12134
0.33	12930	12968	13005	13043	13081	13118	13156	13194	13232	13269
0.36	14058	14095	14132	14169	14207	14244	14281	14319	14356	14393
0.39	15173	15210	15247	15284	15321	15357	15394	15431	15468	15505
0.42	16276	16312	16348	16385	16421	16458	16494	16531	16567	16604
0.45	17364	17400	17436	17472	17508	17544	17580	17616	17652	17688
0.48	18439	18474	18509	18545	18580	18616	18651	18687	18722	18758
0.51	19497	19532	19567	19602	19637	19672	19707	19742	19777	19812
0.54	20540	20574	20609	20643	20678	20712	20746	20781	20815	20850
0.57	21566	21600	21634	21667	21701	21735	21769	21803	21836	21870
0.60	22575	22608	22641	22674	22707	22741	22774	22807	22840	22874
0.63	23565	23598	23630	23663	23695	23728	23761	23793	23826	23859

---

*Adapted from Table IV, C. B. Davenport's *Statistical Methods*.

Table VI—*Continued*

$\frac{x}{\sigma}$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
0.66	24537	24569	24601	24633	24665	24697	24729	24761	24793	24825
0.67	24857	24889	24920	24952	24984	25016	25048	25079	25111	25143
0.69	25490	25521	25553	25584	25615	25647	25678	25709	25741	25772
0.72	26424	26454	26485	26516	26546	26577	26608	26638	26669	26700
0.75	27337	27367	27397	27427	27457	27487	27517	27547	27577	27607
0.78	28230	28260	28289	28318	28347	28377	28406	28435	28465	28494
0.81	29103	29132	29160	29189	29217	29246	29274	29303	29332	29360
0.84	29954	29982	30010	30038	30066	30094	30122	30150	30178	30206
0.87	30785	30812	30839	30866	30894	30921	30948	30975	31002	31030
0.90	31594	31620	31647	31673	31700	31726	31753	31780	31806	31832
0.93	32381	32407	32433	32459	32484	32510	32536	32562	32587	32613
0.96	33147	33172	33197	33222	33247	33272	33297	33322	33347	33373
0.99	33891	33915	33940	33964	33988	34013	34037	34061	34086	34110
1.02	34613	34637	34661	34684	34708	34731	34755	34778	34802	34826
1.05	35314	35337	35360	35382	35405	35428	35451	35474	35497	35520
1.08	35993	—	—	—	—	—	—	—	—	—
—	—	015	037	059	081	103	125	148	170	192
1.11	36650	671	693	714	735	757	778	800	821	843
1.14	37286	306	327	348	368	389	410	430	451	472
1.17	900	920	940	960	980	—	—	—	—	—
—	—	—	—	—	—	000	020	040	060	080
1.20	38493	512	531	551	570	589	609	628	647	667
1.23	39065	084	102	121	139	158	177	195	214	232
1.26	617	634	652	670	688	706	724	742	760	778
1.29	40147	165	182	199	216	233	251	268	285	303
1.32	658	676	692	709	725	742	758	775	792	808
1.35	41149	165	181	197	213	229	245	261	277	292

TABLE VI—*Continued*

$\frac{x}{\sigma}$	0	1	2	3	4	5	6	7	8	9
1.38	41621	637	652	667	683	698	713	728	744	759
1.41	42073	088	102	117	131	146	161	175	190	205
1.44	507	521	535	549	563	577	591	605	619	633
1.47	922	935	949	962	975	989	—	—	—	—
	—	—	—	—	—	—	002	016	029	043
1.50	43319	332	345	358	371	383	396	409	422	435
1.53	699	711	724	736	748	760	773	785	797	810
1.56	44062	074	085	097	109	120	132	144	156	167
1.59	408	419	430	442	453	464	475	486	498	509
1.62	738	749	760	770	781	791	802	813	823	834
1.65	45053	063	073	083	093	103	114	124	134	144
1.68	352	362	371	381	391	400	410	419	429	439
1.71	637	646	655	664	673	682	692	701	710	719
1.74	907	916	924	933	942	950	959	968	977	985
1.77	46164	172	180	188	196	205	213	221	230	238
1.80	407	415	423	430	438	446	454	462	469	477
1.83	638	645	652	660	667	674	682	689	697	704
1.86	856	863	870	877	884	891	898	905	912	919
1.89	47062	069	075	082	088	095	102	108	115	122
1.92	257	263	270	276	282	288	294	301	307	313
1.95	441	447	453	459	465	471	476	482	488	494
1.98	615	620	626	631	637	643	648	654	659	665
2.01	778	784	789	794	799	804	810	815	820	826
2.04	932	937	942	947	952	957	962	967	972	977
2.07	48077	082	087	091	096	100	105	110	114	119
2.10	214	218	222	227	231	235	240	244	248	253
2.13	341	345	350	354	358	362	366	370	374	378

TABLE VI—*Continued*

$\frac{x}{\sigma}$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
2.16	48461	465	469	473	477	480	484	488	492	496
2.19	574	577	581	584	588	592	595	599	602	606
2.22	679	682	686	689	692	696	699	702	706	709
2.25	778	781	784	787	790	793	796	799	803	806
2.28	870	872	875	878	881	884	887	890	893	896
2.31	956	958	961	964	966	969	972	975	977	980
2.34	49036	038	041	043	046	048	051	054	056	059
2.37	111	113	115	118	120	122	125	127	130	132
2.40	180	182	185	187	189	191	193	196	198	200
2.43	245	247	249	251	253	255	257	259	261	264
2.46	305	307	309	311	313	315	317	319	321	323
2.49	361	363	365	367	368	370	372	374	375	377
2.7	653	664	674	683	693	702	711	720	728	736
3.0	865	869	873	878	882	886	889	893	897	900
3.3	952	953	955	957	958	960	961	962	964	965
3.6	984	985	985	986	986	987	987	988	988	989
3.9	995	995	996	996	996	996	996	996	997	997

TABLE VII

TABLE GIVING ODDS FOR VARIOUS VALUES OF $\frac{D}{P.E.}$

$\frac{D}{P.E.}$	ODDS AGAINST SUCH A DIFFERENCE OCCURRING DUE TO CHANCE	$\frac{D}{P.E.}$	ODDS AGAINST SUCH A DIFFERENCE OCCURRING DUE TO CHANCE
1.00	1.00 : 1	3.25	34.24 : 1
1.05	1.09 : 1	3.30	37.40 : 1
1.10	1.18 : 1	3.35	40.95 : 1
1.15	1.28 : 1	3.40	44.79 : 1
1.20	1.39 : 1	3.45	49.10 : 1
1.25	1.51 : 1	3.50	53.82 : 1
1.30	1.63 : 1	3.55	59.10 : 1
1.35	1.76 : 1	3.60	64.88 : 1
1.40	1.90 : 1	3.65	71.36 : 1
1.45	2.05 : 1	3.70	78.49 : 1
1.50	2.21 : 1	3.75	86.57 : 1
1.55	2.38 : 1	3.80	95.34 : 1
1.60	2.57 : 1	3.85	105.38 : 1
1.65	2.76 : 1	3.90	116.37 : 1
1.70	2.98 : 1	3.95	128.53 : 1
1.75	3.20 : 1	4.00	142.27 : 1
1.80	3.45 : 1	4.05	157.73 : 1
1.85	3.71 : 1	4.10	175.06 : 1
1.90	4.00 : 1	4.15	193.55 : 1
1.95	4.31 : 1	4.20	215.45 : 1
2.00	4.64 : 1	4.25	240.55 : 1
2.05	5.00 : 1	4.30	266.38 : 1
2.10	5.38 : 1	4.35	298.40 : 1
2.15	5.80 : 1	4.40	332.33 : 1
2.20	6.25 : 1	4.45	372.13 : 1
2.25	6.74 : 1	4.50	415.67 : 1
2.30	7.28 : 1	4.55	466.29 : 1
2.35	7.85 : 1	4.60	519.83 : 1
2.40	8.48 : 1	4.65	587.24 : 1
2.45	9.16 : 1	4.70	656.89 : 1
2.50	9.90 : 1	4.75	734.29 : 1
2.55	10.70 : 1	4.80	832.33 : 1
2.60	11.58 : 1	4.85	924.93 : 1
2.65	12.54 : 1	4.90	1040.67 : 1
2.70	13.58 : 1	4.95	1189.48 : 1
2.75	14.72 : 1	5.00	1314.79 : 1
2.80	15.97 : 1	5.20	2271.73 : 1
2.85	17.33 : 1	5.40	3570.43 : 1
2.90	18.81 : 1	5.60	6249.00 : 1
2.95	20.45 : 1	5.80	9999.00 : 1
3.00	22.23 : 1	6.00	16665.67 : 1
3.05	24.20 : 1	6.50	49999.00 : 1
3.10	26.37 : 1	7.00	427093.90 : 1
3.15	28.74 : 1	7.50	2368544.71 : 1
3.20	31.36 : 1	8.00	14662755.60 : 1

TABLE VIII

THE CALCULATED ODDS FOR THE *Z* VALUES OF 'STUDENT'S' TABLE FOR ESTIMATING THE PROBABILITY OF THE SIGNIFICANCE OF THE RESULT

Z	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9
.1	1.14	1.22	1.29	1.35	1.40	1.46	1.50	1.55
.15	1.21	1.35	1.46	1.56	1.66	1.75	1.83	1.92
.2	1.29	1.49	1.66	1.82	1.97	2.12	2.26	2.41
.25	1.37	1.64	1.88	2.10	2.32	2.54	2.75	2.97
.3	1.46	1.81	2.13	2.44	2.76	3.08	3.41	3.75
.35	1.54	1.98	2.40	2.81	3.24	3.68	4.14	4.62
.4	1.64	2.18	2.72	3.27	3.85	4.48	5.15	5.88
.45	1.73	2.39	3.05	3.75	4.51	5.33	6.24	7.24
.5	1.84	2.62	3.44	4.35	5.36	6.50	7.80	9.26
.55	1.94	2.85	3.85	4.97	6.25	7.72	9.42	11.4
.6	2.05	3.12	4.33	5.75	7.42	9.42	11.8	14.6
.65	2.16	3.39	4.82	6.54	8.62	11.2	14.2	17.9
.7	2.27	3.69	5.41	7.55	10.2	13.6	17.8	23.1
.75	2.39	3.99	5.99	8.55	11.8	16.0	21.4	28.3
.8	2.51	4.33	6.70	9.82	14.0	19.5	26.8	36.5
.85	2.62	4.66	7.39	11.1	16.1	22.9	32.1	44.5
.9	2.75	5.04	8.22	12.7	18.9	27.7	40.0	57.1
.95	2.87	5.41	9.03	14.2	21.7	32.4	47.8	69.4
1.0	3.00	5.83	10.0	16.2	25.5	39.2	59.2	89.1
1.05	3.12	6.24	11.0	18.2	29.1	45.7	70.4	108.
1.1	3.26	6.69	12.1	20.6	34.0	54.9	87.5	138.
1.15	3.39	7.13	13.2	22.9	38.7	63.5	103.	166.
1.2	3.52	7.63	14.5	25.9	44.9	75.9	127.	212.
1.25	3.65	8.11	15.7	28.8	50.8	87.5	151.	255.
1.3	3.79	8.64	17.2	32.3	58.5	104.	184.	322.
1.35	3.92	9.16	18.6	35.8	66.1	119.	216.	384.
1.4	4.07	9.74	20.3	40.0	75.9	142.	262.	475.
1.45	4.20	10.3	21.9	44.0	85.2	163.	302.	555.
1.5	4.34	10.9	23.9	49.0	98.0	191.	369.	713.
1.55	4.48	11.5	25.7	53.9	109.	216.	434.	832.
1.6	4.62	12.2	27.7	60.0	124.	255.	525.	999.
1.65	4.76	12.8	29.8	65.7	138.	285.	587.	1110.
1.7	4.91	13.5	32.2	72.5	158.	332.	713.	1428.
1.75	5.05	14.2	34.5	79.0	174.	369.	832.	1666.
1.8	5.20	14.9	37.0	86.7	199.	434.	999.	1999.
1.85	5.34	15.6	39.5	94.2	216.	499.	1110.	2499.
1.9	5.49	16.4	42.5	103.	243.	587.	1249.	3332.
1.95	5.63	17.1	45.1	111.	269.	624.	1428.	3332.
2.0	5.78	17.9	48.3	122.	302.	713.	1666.	4999.
2.05	5.92	18.7	51.4	132.	332.	768.	1999.	4999.
2.1	6.07	19.6	54.9	144.	369.	908.	2499.	4999.
2.15	6.21	20.4	58.2	155.	399.	999.	2499.	4999.
2.2	6.36	21.3	61.9	168.	454.	1249.	3332.	9999.
2.25	6.51	22.1	65.2	181.	499.	1249.	3332.	
2.3	6.66	23.1	69.4	199.	555.	1428.	4999.	
2.35	6.81	24.0	73.1	212.	587.	1666.	4999.	
2.4	6.96	25.0	77.7	232.	666.	1999.	4999.	
2.45	7.10	25.9	81.6	249.	713.	1999.	4999.	
2.5	7.26	26.9	86.7	269.	768.	2499.	4999.	
2.55	7.40	27.9	91.6	285.	832.	2499.	4999.	
2.6	7.55	29.0	97.0	302.	908.	2499.	9999.	
2.65	7.70	30.1	102.	322.	999.	2499.		
2.7	7.86	31.2	108.	356.	1110.	3332.		

TABLE VIII—*Continued*

Z	n=10	n=11	n=12	n=13	n=14	n=15	n=16	n=17
.1	1.59	1.64	1.68	1.72	1.76	1.80	1.84	1.88
.15	2.00	2.08	2.16	2.24	2.31	2.39	2.47	2.54
.2	2.55	2.70	2.84	2.99	3.14	3.29	3.44	3.60
.25	3.19	3.41	3.64	3.87	4.11	4.36	4.60	4.86
.3	4.11	4.48	4.86	5.27	5.69	6.13	6.59	7.08
.35	5.13	5.67	6.24	6.84	7.47	8.15	8.86	9.62
.4	6.67	7.53	8.45	9.47	10.6	11.8	13.1	14.5
.45	8.34	9.54	10.9	12.3	13.9	15.7	17.7	19.8
.5	10.9	12.8	14.9	17.3	20.1	23.3	26.8	30.8
.55	13.6	16.2	19.2	22.7	26.7	31.3	36.5	42.7
.6	18.0	22.0	26.8	32.4	39.3	47.3	56.8	68.4
.65	22.5	27.9	34.6	42.5	52.2	63.9	77.7	95.2
.7	29.8	38.1	48.5	61.5	77.7	99.0	124.	155.
.75	37.2	48.3	62.7	81.0	104.	134.	171.	216.
.8	49.3	66.1	88.3	118.	158.	207.	277.	356.
.85	61.1	83.7	114.	155.	207.	277.	369.	499.
.9	81.0	114.	160.	226.	311.	434.	587.	832.
.95	100.	144.	207.	293.	416.	587.	832.	1110.
1.0	132.	195.	293.	434.	624.	908.	1428.	1999.
1.05	163.	243.	369.	555.	832.	1249.	1999.	2499.
1.1	216.	332.	525.	832.	1249.	1999.	3332.	4999.
1.15	262.	416.	666.	999.	1666.	2499.	3332.	4999.
1.2	344.	555.	908.	1428.	2499.	3332.	4999.	9999.
1.25	416.	713.	1110.	1666.	3332.	4999.	4999.	
1.3	555.	999.	1666.	2499.	4999.	9999.	9999.	
1.35	666.	1249.	1999.	3332.	4999.			
1.4	908.	1666.	3332.	4999.	9999.			
1.45	1110.	1999.	3332.	4999.				
1.5	1428.	2499.	4999.	9999.				
1.55	1666.	2499.	4999.					
1.6	1999.	3332.	9999.					
1.65	2499.	3332.						
1.7	3332.	4999.						
1.75	3332.	4999.						
1.8	4999.	9999.						
1.85	4999.							
1.9	9999.							

TABLE VIII—*Continued*

Z	n=2	n=3	n=4	n=5	n=6	n=7
2.75	8.00	32.2	113.	369.	1110.	3332.
2.8	8.16	33.4	118.	399.	1249.	4999.
2.85	8.30	34.5	124.	416.	1249.	4999.
2.9	8.46	35.6	131.	454.	1428.	4999.
2.95	8.61	36.7	136.	475.	1428.	4999.
3.0	8.77	37.9	144.	525.	1666.	4999.
3.05	8.91	39.2	151.	555.	1666.	4999.
3.1	9.07	40.5	158.	587.	1999.	9999.
3.15	9.21	41.6	163.	587.	1999.	
3.2	9.37	42.9	171.	624.	2499.	
3.25	9.52	44.0	178.	666.	2499.	
3.3	9.67	45.5	188.	713.	2499.	
3.35	9.82	46.8	195.	768.	2499.	
3.4	9.98	48.3	203.	832.	3332.	
3.45	10.1	49.5	212.	832.	3332.	
3.5	10.3	51.1	221.	908.	3332.	
4.0	11.8	66.1	322.	1666.	4999.	
4.5	13.3	83.0	434.	2499.	4999.	
5.0	14.9	102.	624.	3332.	9999.	
5.5	16.5	122.	832.	4999.		
6.0	18.0	146.	999.	9999.		
6.5	19.6	171.	1249.			
7.0	21.1	199.	1666.			
7.5	22.7	226.	1999.			
8.0	24.3	255.	2499.			
8.5	25.8	293.	2499.			
9.0	27.4	322.	3332.			
9.5	28.9	369.	3332.			
10.0	30.5	399.	4999.			
15.0	46.2	908.	9999.			
20.0	61.9	1666.				
25.0	77.7	2499.				
30.0	93.3	3332.				
35.0	109.	4999.				
40.0	124.	4999.				
45.0	140.	9999.				
50.0	153.					
60.0	188.					
70.0	221.					
80.0	249.					
90.0	285.					
100.0	311.					
120.0	369.					
140.0	434.					
150.0	475.					
160.0	499.					
180.0	555.					
200.0	624.					
250.0	768.					
300.0	908.					
350.0	1110.					
400.0	1249.					
450.0	1428.					
500.0	1666.					
600.0	1999.					
700.0	1999.					
1000.0	3332.					
1500.0	4999.					
2000.0	4999.					
3000.0	9999.					

Table VIII—*Continued*

Z	n=18	n=19	n=20	n=21	n=22	n=23	n=24	n=25	n=26	n=27	n=28	n=29	n=30
.1	1.92	1.96	2.00	2.03	2.07	2.11	2.15	2.18	2.22	2.26	2.29	2.33	2.37
.15	2.62	2.69	2.77	2.84	2.92	2.99	3.07	3.14	3.22	3.30	3.37	3.45	3.52
.2	3.75	3.91	4.08	4.24	4.41	4.58	4.76	4.94	5.12	5.31	5.50	5.69	5.90
.25	5.12	5.39	5.66	5.94	6.23	6.52	6.83	7.14	7.46	7.79	8.12	8.47	8.82
.3	7.59	8.12	8.68	9.27	9.89	10.5	11.2	11.9	12.7	13.5	14.3	15.2	16.1
.35	10.4	11.3	12.2	13.1	14.1	15.2	16.3	17.5	18.7	20.1	21.5	23.0	24.5
.4	16.0	17.7	19.5	21.5	23.7	26.0	28.6	31.4	34.3	37.8	41.2	45.3	49.3
.45	22.2	24.8	27.7	30.7	34.2	38.1	42.1	46.6	51.6	57.1	63.1	69.9	76.5
.5	35.4	40.7	46.4	53.1	60.7	69.4	79.0	89.9	102.	117.	132.	151.	171.
.55	49.5	57.5	66.6	77.1	89.1	103.	118.	136.	155.	181.	207.	237.	269.
.6	81.6	98.0	117.	140.	166.	199.	237.	277.	332.	399.	475.	555.	666.
.65	115.	140.	168.	203.	243.	302.	356.	434.	525.	624.	768.	908.	1110.
.7	195.	249.	302.	384.	475.	624.	768.	999.	1249.	1666.	1999.	2499.	3332.
.75	277.	356.	434.	555.	713.	908.	1110.	1428.	1999.	2499.	3332.	3332.	4999.
.8	475.	624.	832.	1110.	1428.	1999.	2499.	3332.	4999.	9999.	9999.	9999.	9999.
.85	666.	908.	1110.	1428.	1999.	2499.	3332.	4999.	4999.				
.9	1110.	1666.	1999.	2499.	3332.	4999.	9999.	9999.	9999.				
.95	1666.	2499.	3332.	3332.	4999.	4999.							
1.0	3332.	4999.	9999.	9999.	9999.	9999.							
1.05	4999.	4999.											
1.1	9999.	9999.											

Table X

Values of F and t

		Degrees of Freedom for Greater Mean Square										
		1	2	3	4	5	6	8	12	24	∞	Values of t
Degrees of Freedom for Smaller Mean Square	1	161.45 **4052.10**	199.50 **4999.03**	215.72 **5403.49**	224.57 **5625.14**	230.17 **5764.08**	233.97 **5859.39**	238.89 **5981.34**	243.91 **6105.83**	249.04 **6234.16**	254.32 **6366.48**	12.706 **63.657**
	2	18.51 **98.49**	19.00 **99.01**	19.16 **99.17**	19.25 **99.25**	19.30 **99.30**	19.33 **99.33**	19.37 **99.36**	19.41 **99.42**	19.45 **99.46**	19.50 **99.50**	4.303 **9.925**
	3	10.13 **34.12**	9.55 **30.81**	9.28 **29.46**	9.12 **28.71**	9.01 **28.24**	8.94 **27.91**	8.84 **27.49**	8.74 **27.05**	8.64 **26.60**	8.53 **26.12**	3.182 **5.841**
	4	7.71 **21.20**	6.94 **18.00**	6.59 **16.69**	6.39 **15.98**	6.26 **15.52**	6.16 **15.21**	6.04 **14.80**	5.91 **14.37**	5.77 **13.93**	5.63 **13.46**	2.776 **4.604**
	5	6.61 **16.26**	5.79 **13.27**	5.41 **12.06**	5.19 **11.39**	5.05 **10.97**	4.95 **10.67**	4.82 **10.27**	4.68 **9.89**	4.53 **9.47**	4.36 **9.02**	2.571 **4.032**
	6	5.99 **13.74**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**	4.39 **8.75**	4.28 **8.47**	4.15 **8.10**	4.00 **7.72**	3.84 **7.31**	3.67 **6.88**	2.447 **3.707**
	7	5.59 **12.25**	4.74 **9.55**	4.35 **8.45**	4.12 **7.85**	3.97 **7.46**	3.87 **7.19**	3.73 **6.84**	3.57 **6.47**	3.41 **6.07**	3.23 **5.65**	2.365 **3.499**
	8	5.32 **11.26**	4.46 **8.65**	4.07 **7.59**	3.84 **7.01**	3.69 **6.63**	3.58 **6.37**	3.44 **6.03**	3.28 **5.67**	3.12 **5.28**	2.93 **4.86**	2.306 **3.355**
	9	5.12 **10.56**	4.26 **8.02**	3.86 **6.99**	3.63 **6.42**	3.48 **6.06**	3.37 **5.80**	3.23 **5.47**	3.07 **5.11**	2.90 **4.73**	2.71 **4.31**	2.262 **3.250**
	10	4.96 **10.04**	4.10 **7.56**	3.71 **6.55**	3.48 **5.99**	3.33 **5.64**	3.22 **5.39**	3.07 **5.06**	2.91 **4.71**	2.74 **4.33**	2.54 **3.91**	2.228 **3.169**
	11	4.84 **9.65**	3.98 **7.20**	3.59 **6.22**	3.36 **5.67**	3.20 **5.32**	3.09 **5.07**	2.95 **4.74**	2.79 **4.40**	2.61 **4.02**	2.40 **3.60**	2.201 **3.106**

TABLE X—*Continued*

		Degrees of Freedom for Greater Mean Square										
		1	2	3	4	5	6	8	12	24	∞	Values of *t*
Degrees of Freedom for Smaller Mean Square	12	4.75 **9.33**	3.88 **6.93**	3.49 **5.95**	3.26 **5.41**	3.11 **5.06**	3.00 **4.82**	2.85 **4.50**	2.69 **4.16**	2.50 **3.78**	2.30 **3.36**	2.179 **3.055**
	13	4.67 **9.07**	3.80 **6.70**	3.41 **5.74**	3.18 **5.20**	3.02 **4.86**	2.92 **4.62**	2.77 **4.30**	2.60 **3.96**	2.42 **3.59**	2.21 **3.16**	2.160 **3.012**
	14	4.60 **8.86**	3.74 **6.51**	3.34 **5.56**	3.11 **5.03**	2.96 **4.69**	2.85 **4.46**	2.70 **4.14**	2.53 **3.80**	2.35 **3.43**	2.13 **3.00**	2.145 **2.977**
	15	4.54 **8.68**	3.68 **6.36**	3.29 **5.42**	3.06 **4.89**	2.90 **4.56**	2.79 **4.32**	2.64 **4.00**	2.48 **3.67**	2.29 **3.29**	2.07 **2.87**	2.131 **2.947**
	16	4.49 **8.53**	3.63 **6.23**	3.24 **5.29**	3.01 **4.77**	2.85 **4.44**	2.74 **4.20**	2.59 **3.89**	2.42 **3.55**	2.24 **3.18**	2.01 **2.75**	2.120 **2.921**
	17	4.45 **8.40**	3.59 **6.11**	3.20 **5.18**	2.96 **4.67**	2.81 **4.34**	2.70 **4.10**	2.55 **3.79**	2.38 **3.45**	2.19 **3.08**	1.96 **2.65**	2.110 **2.898**
	18	4.41 **8.28**	3.55 **6.01**	3.16 **5.09**	2.93 **4.58**	2.77 **4.25**	2.66 **4.01**	2.51 **3.71**	2.34 **3.37**	2.15 **3.01**	1.92 **2.57**	2.101 **2.878**
	19	4.38 **8.18**	3.52 **5.93**	3.13 **5.01**	2.90 **4.50**	2.74 **4.17**	2.63 **3.94**	2.48 **3.63**	2.31 **3.30**	2.11 **2.92**	1.88 **2.49**	2.093 **2.861**
	20	4.35 **8.10**	3.49 **5.85**	3.10 **4.94**	2.87 **4.43**	2.71 **4.10**	2.60 **3.87**	2.45 **3.56**	2.28 **3.23**	2.08 **2.86**	1.84 **2.42**	2.086 **2.845**
	21	4.32 **8.02**	3.47 **5.78**	3.07 **4.87**	2.84 **4.37**	2.68 **4.04**	2.57 **3.81**	2.42 **3.51**	2.25 **3.17**	2.05 **2.80**	1.81 **2.36**	2.080 **2.831**
	22	4.30 **7.94**	3.44 **5.72**	3.05 **4.82**	2.82 **4.31**	2.66 **3.99**	2.55 **3.75**	2.40 **3.45**	2.23 **3.12**	2.03 **2.75**	1.78 **2.30**	2.074 **2.819**
	23	4.28 **7.88**	3.42 **5.66**	3.03 **4.76**	2.80 **4.26**	2.64 **3.94**	2.53 **3.71**	2.38 **3.41**	2.20 **3.07**	2.00 **2.70**	1.76 **2.26**	2.069 **2.807**

TABLE X—*Continued*

DEGREES OF FREEDOM FOR SMALLER MEAN SQUARE	DEGREES OF FREEDOM FOR GREATER MEAN SQUARE										
	1	2	3	4	5	6	8	12	24	∞	VALUES OF t
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73	2.064
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.36**	**3.03**	**2.66**	**2.21**	**2.797**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71	2.060
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.32**	**2.99**	**2.62**	**2.17**	**2.787**
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69	2.056
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.29**	**2.96**	**2.58**	**2.13**	**2.779**
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67	2.052
	7.68	**5.49**	**4.60**	**4.11**	**3.78**	**3.56**	**3.26**	**2.93**	**2.55**	**2.10**	**2.771**
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65	2.048
	7.64	**5.45**	**4.57**	**4.07**	**3.75**	**3.53**	**3.23**	**2.90**	**2.52**	**2.06**	**2.763**
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64	2.045
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.20**	**2.87**	**2.49**	**2.03**	**2.756**
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62	2.042
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.17**	**2.84**	**2.47**	**2.01**	**2.750**
35	4.12	3.26	2.87	2.64	2.48	2.37	2.22	2.04	1.83	1.57	2.030
	7.42	**5.27**	**4.40**	**3.91**	**3.59**	**3.37**	**3.07**	**2.74**	**2.37**	**1.90**	**2.724**
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.52	2.021
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**2.99**	**2.66**	**2.29**	**1.82**	**2.704**
45	4.06	3.21	2.81	2.58	2.42	2.31	2.15	1.97	1.76	1.48	2.014
	7.23	**5.11**	**4.25**	**3.77**	**3.45**	**3.23**	**2.94**	**2.61**	**2.23**	**1.75**	**2.690**
50	4.03	3.18	2.79	2.56	2.40	2.29	2.13	1.95	1.74	1.44	2.008
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.19**	**2.89**	**2.56**	**2.18**	**1.68**	**2.678**
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39	2.000
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.82**	**2.50**	**2.12**	**1.60**	**2.660**

TABLE X—*Continued*

		Degrees of Freedom for Greater Mean Square									
		1	2	3	4	5	6	8	12	24	∞	Values of t
Degrees of Freedom for Smaller Mean Square	70	3.98 **7.01**	3.13 **4.92**	2.74 **4.07**	2.50 **3.60**	2.35 **3.29**	2.23 **3.07**	2.07 **2.78**	1.89 **2.45**	1.67 **2.07**	1.35 **1.53**	1.994 **2.648**
	80	3.96 **6.96**	3.11 **4.88**	2.72 **4.04**	2.49 **3.56**	2.33 **3.26**	2.21 **3.04**	2.06 **2.74**	1.88 **2.42**	1.65 **2.03**	1.31 **1.47**	1.990 **2.638**
	90	3.95 **6.92**	3.10 **4.85**	2.71 **4.01**	2.47 **3.53**	2.32 **3.23**	2.20 **3.01**	2.04 **2.72**	1.86 **2.39**	1.64 **2.00**	1.28 **1.43**	1.987 **2.632**
	100	3.94 **6.90**	3.09 **4.82**	2.70 **3.98**	2.46 **3.51**	2.30 **3.21**	2.19 **2.99**	2.03 **2.69**	1.85 **2.37**	1.63 **1.98**	1.26 **1.39**	1.984 **2.626**
	125	3.92 **6.84**	3.07 **4.78**	2.68 **3.94**	2.44 **3.47**	2.29 **3.17**	2.17 **2.95**	2.01 **2.66**	1.83 **2.33**	1.60 **1.94**	1.21 **1.32**	1.979 **2.616**
	150	3.90 **6.81**	3.06 **4.75**	2.66 **3.91**	2.43 **3.45**	2.27 **3.14**	2.16 **2.92**	2.00 **2.63**	1.82 **2.31**	1.59 **1.92**	1.18 **1.27**	1.976 **2.609**
	200	3.89 **6.76**	3.04 **4.71**	2.65 **3.88**	2.42 **3.41**	2.26 **3.11**	2.14 **2.89**	1.98 **2.60**	1.80 **2.28**	1.57 **1.88**	1.14 **1.21**	1.972 **2.601**
	300	3.87 **6.72**	3.03 **4.68**	2.64 **3.85**	2.41 **3.38**	2.25 **3.08**	2.13 **2.86**	1.97 **2.57**	1.79 **2.24**	1.55 **1.85**	1.10 **1.14**	1.968 **2.592**
	400	3.86 **6.70**	3.02 **4.66**	2.63 **3.83**	2.40 **3.37**	2.24 **3.06**	2.12 **2.85**	1.96 **2.56**	1.78 **2.23**	1.54 **1.84**	1.07 **1.11**	1.966 **2.588**
	500	3.86 **6.69**	3.01 **4.65**	2.62 **3.82**	2.39 **3.36**	2.23 **3.65**	2.11 **2.84**	1.96 **2.55**	1.77 **2.22**	1.54 **1.83**	1.06 **1.08**	1.965 **2.586**
	1000	3.85 **6.66**	3.00 **4.63**	2.61 **3.80**	2.38 **3.34**	2.22 **3.04**	2.10 **2.82**	1.95 **2.53**	1.76 **2.20**	1.53 **1.81**	1.03 **1.04**	1.962 **2.581**
	∞	3.84 **6.64**	2.99 **4.60**	2.60 **3.78**	2.37 **3.32**	2.21 **3.02**	2.09 **2.80**	1.94 **2.51**	1.75 **2.18**	1.52 **1.79**		1.960 **2.576**

TABLE XI

SIGNIFICANT VALUES OF r AND R

DEGREES OF FREEDOM	NUMBER OF VARIABLES								
	2	3	4	5	6	7	9	13	25
1	.997 **1.000**	.999 **1.000**	.999 **1.000**	.999 **1.000**	1.000 **1.000**	1.000 **1.000**	1.000 **1.000**	1.000 **1.000**	1.000 **1.000**
2	.950 **.990**	.975 **.995**	.983 **.997**	.987 **.998**	.990 **.998**	.992 **.998**	.994 **.999**	.996 **.999**	.998 **1.000**
3	.878 **.959**	.930 **.976**	.950 **.983**	.961 **.987**	.968 **.990**	.973 **.991**	.979 **.993**	.986 **.995**	.993 **.998**
4	.811 **.917**	.881 **.949**	.912 **.962**	.930 **.970**	.942 **.975**	.950 **.979**	.961 **.984**	.973 **.989**	.986 **.994**
5	.754 **.874**	.836 **.917**	.874 **.937**	.898 **.949**	.914 **.957**	.925 **.963**	.941 **.971**	.958 **.980**	.978 **.989**
6	.707 **.834**	.795 **.886**	.839 **.911**	.867 **.927**	.886 **.938**	.900 **.946**	.920 **.957**	.943 **.969**	.969 **.983**
7	.666 **.798**	.758 **.855**	.807 **.885**	.838 **.904**	.860 **.918**	.876 **.928**	.900 **.942**	.927 **.958**	.960 **.977**
8	.632 **.765**	.726 **.827**	.777 **.860**	.811 **.882**	.835 **.898**	.854 **.909**	.880 **.926**	.912 **.946**	.950 **.970**
9	.602 **.735**	.697 **.800**	.750 **.836**	.786 **.861**	.812 **.878**	.832 **.891**	.861 **.911**	.897 **.934**	.941 **.963**
10	.576 **.708**	.671 **.776**	.726 **.814**	.763 **.840**	.790 **.859**	.812 **.874**	.843 **.895**	.882 **.922**	.932 **.955**
11	.553 **.684**	.648 **.753**	.703 **.793**	.741 **.821**	.770 **.841**	.792 **.857**	.826 **.880**	.868 **.910**	.922 **.948**
12	.532 **.661**	.627 **.732**	.683 **.773**	.722 **.802**	.751 **.824**	.774 **.841**	.809 **.866**	.854 **.898**	.913 **.940**
13	.514 **.641**	.608 **.712**	.664 **.755**	.703 **.785**	.733 **.807**	.757 **.825**	.794 **.852**	.840 **.886**	.904 **.932**
14	.497 **.623**	.590 **.694**	.646 **.737**	.686 **.768**	.717 **.792**	.741 **.810**	.779 **.838**	.828 **.875**	.895 **.924**
15	.482 **.606**	.574 **.677**	.630 **.721**	.670 **.752**	.701 **.776**	.726 **.796**	.765 **.825**	.815 **.864**	.886 **.917**
16	.468 **.590**	.559 **.662**	.615 **.706**	.655 **.738**	.686 **.762**	.712 **.782**	.751 **.813**	.803 **.853**	.878 **.909**
17	.456 **.575**	.545 **.647**	.601 **.691**	.641 **.724**	.673 **.749**	.698 **.769**	.738 **.800**	.792 **.842**	.869 **.902**
18	.444 **.561**	.532 **.633**	.587 **.678**	.628 **.710**	.660 **.736**	.686 **.756**	.726 **.789**	.781 **.832**	.861 **.894**

TABLE XI—*Continued*

DEGREES OF FREEDOM	NUMBER OF VARIABLES								
	2	3	4	5	6	7	9	13	25
19	.433 **.549**	.520 **.620**	.575 **.665**	.615 **.698**	.647 **.723**	.674 **.744**	.714 **.778**	.770 **.822**	.853 **.887**
20	.423 **.537**	.509 **.608**	.563 **.652**	.604 **.685**	.636 **.712**	.662 **.733**	.703 **.767**	.760 **.812**	.845 **.880**
21	.413 **.526**	.498 **.596**	.552 **.641**	.592 **.674**	.624 **.700**	.651 **.722**	.693 **.756**	.750 **.803**	.837 **.873**
22	.404 **.515**	.488 **.585**	.542 **.630**	.582 **.663**	.614 **.690**	.640 **.712**	.682 **.746**	.740 **.794**	.830 **.866**
23	.396 **.505**	.479 **.574**	.532 **.619**	.572 **.652**	.604 **.679**	.630 **.701**	.673 **.736**	.731 **.785**	.823 **.859**
24	.388 **.496**	.470 **.565**	.523 **.609**	.562 **.642**	.594 **.669**	.621 **.692**	.663 **.727**	.722 **.776**	.815 **.852**
25	.381 **.487**	.462 **.555**	.514 **.600**	.553 **.633**	.585 **.660**	.612 **.682**	.654 **.718**	.714 **.763**	.808 **.846**
26	.374 **.478**	.454 **.546**	.506 **.590**	.545 **.624**	.576 **.651**	.603 **.673**	.645 **.709**	.706 **.760**	.802 **.839**
27	.367 **.470**	.446 **.538**	.498 **.582**	.536 **.615**	.568 **.642**	.594 **.664**	.637 **.701**	.698 **.752**	.795 **.833**
28	.361 **.463**	.439 **.530**	.490 **.573**	.529 **.606**	.560 **.634**	.586 **.656**	.629 **.692**	.690 **.744**	.788 **.827**
29	.355 **.456**	.432 **.522**	.482 **.565**	.521 **.598**	.552 **.625**	.579 **.648**	.621 **.685**	.682 **.737**	.782 **.821**
30	.349 **.449**	.426 **.514**	.476 **.558**	.514 **.591**	.545 **.618**	.571 **.640**	.614 **.677**	.675 **.729**	.776 **.815**
35	.325 **.418**	.397 **.481**	.445 **.523**	.482 **.556**	.512 **.582**	.538 **.605**	.580 **.642**	.642 **.696**	.746 **.786**
40	.304 **.393**	.373 **.454**	.419 **.494**	.455 **.526**	.484 **.552**	.509 **.575**	.551 **.612**	.613 **.667**	.720 **.761**
45	.288 **.372**	.353 **.430**	.397 **.470**	.432 **.501**	.460 **.527**	.485 **.549**	.526 **.586**	.587 **.640**	.696 **.737**
50	.273 **.354**	.336 **.410**	.379 **.449**	.412 **.479**	.440 **.504**	.464 **.526**	.504 **.562**	.565 **.617**	.674 **.715**
60	.250 **.325**	.308 **.377**	.348 **.414**	.380 **.442**	.406 **.466**	.429 **.438**	.467 **.523**	.526 **.577**	.636 **.677**
70	.232 **.302**	.286 **.351**	.324 **.385**	.354 **.413**	.379 **.436**	.401 **.456**	.438 **.491**	.495 **.544**	.604 **.644**
80	.217 **.283**	.269 **.330**	.304 **.362**	.332 **.389**	.356 **.411**	.377 **.431**	.413 **.464**	.469 **.516**	.576 **.615**

TABLE XI—*Continued*

DEGREES OF FREEDOM	NUMBER OF VARIABLES								
	2	3	4	5	6	7	9	13	25
90	.205 **.267**	.254 **.312**	.288 **.343**	.315 **.368**	.338 **.390**	.358 **.409**	.392 **.441**	.446 **.492**	.552 **.590**
100	.195 **.254**	.241 **.297**	.274 **.327**	.300 **.351**	.322 **.372**	.341 **.390**	.374 **.421**	.426 **.470**	.530 **.568**
125	.174 **.228**	.216 **.266**	.246 **.294**	.269 **.316**	.290 **.335**	.307 **.352**	.338 **.381**	.387 **.428**	.485 **.521**
150	.159 **.208**	.198 **.244**	.225 **.270**	.247 **.290**	.266 **.308**	.282 **.324**	.310 **.351**	.356 **.395**	.450 **.484**
200	.138 **.181**	.172 **.212**	.196 **.234**	.215 **.253**	.231 **.269**	.246 **.283**	.271 **.307**	.312 **.347**	.398 **.430**
300	.113 **.148**	.141 **.174**	.160 **.192**	.176 **.208**	.190 **.221**	.202 **.233**	.223 **.253**	.258 **.287**	.332 **.359**
400	.098 **.128**	.122 **.151**	.139 **.167**	.153 **.180**	.165 **.192**	.176 **.202**	.194 **.220**	.225 **.250**	.291 **.315**
500	.088 **.115**	.109 **.135**	.124 **.150**	.137 **.162**	.148 **.172**	.157 **.182**	.174 **.198**	.202 **.225**	.262 **.284**
1000	.062 **.081**	.077 **.096**	.088 **.106**	.097 **.115**	.105 **.122**	.112 **.129**	.124 **.141**	.144 **.160**	.188 **.204**

PUBLICATIONS REFERRED TO IN THE TEXT

PUBLICATIONS REFERRED TO IN THE TEXT

ALLAN, F. E., AND WISHART, J., A Method of Estimating the Yield of a Missing Plot in Field Experimental Work. *Journal of Agricultural Science*, Vol. 20. 1930.

CHADDOCK, R. E., *Principles and Methods of Statistics.* Houghton Mifflin, Boston. 1925.

DAY, JAMES W., The Relation of Size, Shape, and Number of Replications of Plats to Probable Error in Field Experimentation. *Journal of the American Society of Agronomy*, Vol. 12. 1920.

ELDERTON, W. PALIN, Tables for Testing the Goodness of Fit of Theory to Observation. *Biometrika*, Vol. I. 1901.

EZEKIEL, MORDECAI, *Methods of Correlation Analysis.* Wiley, New York. 1930.

FISHER, R. A., *Statistical Methods for Research Workers*, 5th Edition. Oliver and Boyd, Edinburgh. 1934.

GARBER, R. J., MCILVAINE, T. C., AND HOOVER, M. M., A Method of Laying Out Experiment Plats. *Journal of the American Society of Agronomy*, Vol. 23. 1931.

HALL, A. D., AND RUSSELL, E. J., Field Trials and Their Interpretation, *Journal of the Board of Agriculture Supplement 7.* Board of Agriculture and Fisheries, London. 1911.

HARRIS, J. ARTHUR, On a Criterion of Substratum Homogeneity (or Heterogeneity) in Field Experiments. *American Naturalist*, Vol. 49. 1915.

— Practical Universality of Field Heterogeneity as a Factor Influencing Plot Yields. *Journal of Agricultural Research*, Vol. XIX. 1920.

— AND SCOFIELD, C. S., Permanence of Differences in the Plots of An Experimental Field. *Journal of Agricultural Research*, Vol. XX. 1920.

HAYES, H. K., Control of Soil Heterogeneity and Use of the Probable Error Concept in Plant Breeding Studies. University of Minnesota Agricultural Experiment Station. *Technical Bul. 30.* 1925.

Illinois Agricultural Experiment Station. Data cited by HAYES, H. K., AND GARBER, R. J. *Breeding Crop Plants*, 2d Edition. McGraw Hill, New York, 1927.

LIVERMORE, J. R., The Interrelations of Various Probability Tables and a Modification of Student's Probability Table for the Argument "t". *Journal of the American Society of Agronomy*, Vol. 26. 1934.

MCCLELLAND, C. K., Some Determinations of Plat Variability. *Journal of the American Society of Agronomy*, Vol. 18. 1926.

MARTIN, JOHN H., Factors Influencing Results from Rate- and Date-of-Seeding Experiments with Wheat in the Western United States. *Journal of the American Society of Agronomy*, Vol. 18. 1926.

MILLS, FREDERICK C., *Statistical Methods.* Holt, New York. 1924.

PARKER, E. R., AND BATCHELOR, L. D., Variation in the Yields of Fruit Trees in Relation to the Planning of Future Experiments. *Hilgardia*, Vol. 7. 1932.

PEARSON, K., On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling. *Philosophical Magazine*, Vol. L, 5th Series. 1900.

— On the Theory of Contingency and Its Relation to Association and Normal Correlation. *Drapers' Company Research Memoirs Biometric Series I.* 1904.

— On the General Theory of Skew Correlation and Non-Linear Regression. *Drapers' Company Research Memoirs Biometric Series II.* 1905.

— On the Influence of Past Experience on Future Expectation. *Philosophical Magazine*, Vol. XIII, 6th Series. 1907.

— On the Correction Necessary for the Correlation Ratio η. *Biometrika*, Vol. XIV. 1923.

— (Editor) *Tables for Statisticians and Biometricians*, 2d Edition. Cambridge University Press, London. 1924.

— On a New Method of Determining "Goodness of Fit." *Biometrika*, Vol. XXVI. 1934.

RICHEY, FREDERICK D., Adjusting Yields to Their Regression on a Moving Average, as a Means of Correcting for Soil Heterogeneity. *Journal of Agricultural Research*, Vol. XXVII. 1924.

— The Moving Average as a Basis for Measuring Correlated Variation in Agronomic Experiments. *Journal of Agricultural Research*, Vol. XXXII. 1926.

SANDERS, H. G., A Note on the Value of Uniformity Trials for Subsequent Experiments. *Journal of Agricultural Science*, Vol. XX. 1930.

SECRIST, HORACE, *An Introduction to Statistical Methods.* Macmillan, New York. 1917.

SHEN, T. H., Field Technic for Determining Comparative Yields in Wheat Under Different Environmental Conditions in China. *Journal of the American Society of Agronomy,* Vol. 22. 1930.

SNEDECOR, GEORGE W., *Calculation and Interpretation of Analysis of Variance and Covariance.* Collegiate Press, Ames. 1934.

SPEARMAN, C., The Proof and Measurement of Association Between Two Things. *American Journal of Psychology,* Vol. 15. 1904.

— A "Foot Rule" for Measuring Correlation. *British Journal of Psychology,* Vol. 2. 1906.

STADLER, L. J., Experiments in Field Plot Technic for the Preliminary Determination of Comparative Yields in the Small Grains. University of Missouri Agricultural Experiment Station. *Research Bul. 49.* 1921.

STRINGFIELD, G. H., Intervarietal Competition Among Small Grains. *Journal of the American Society of Agronomy,* Vol. 19. 1927.

'STUDENT,' The Probable Error of a Mean. *Biometrika,* Vol. VI. 1908. (A More Extended Table.) *Biometrika,* Vol. XI. 1917.

— On Testing Varieties of Cereals. *Biometrika,* Vol, XV. 1923.

— New Tables for Testing the Significance of Observations. *Metron,* Vol. 5. 1925.

— Mathematics and Agronomy. *Journal of the American Society of Agronomy,* Vol. 18. 1926.

TIPPETT, L. H. C., *The Methods of Statistics.* Williams and Norgate, London. 1931.

WALLACE, H. A., AND SNEDECOR, GEORGE W., *Correlation and Machine Calculation,* Revised Edition. *Iowa State College of Agriculture and Mechanic Arts,* Vol. XXX, No. 4. 1931.

YATES, F., The Analysis of Replicated Experiments When the Field Results are Incomplete. *Empire Journal of Experimental Agriculture,* Vol. I. 1933.

YULE, G. UDNY, *An Introduction to the Theory of Statistics,* 6th Edition Griffin, London. 1922.

INDEX

Allan, F. E., 364, 365
Arithmetic mean, definition of, 40; methods of determining, 41; treated algebraically, 49. *See* Averages
Array, 100
Average deviation, calculation of, 67; definition of, 67; use of, 74
Averages, comparison of qualities and usefulness of, 60, 64; necessary qualifications of, 40; weighting, 51. *See* Arithmetic mean, Geometric mean, Median, Mode, Moving average

Bar diagram, 28
Batchelor, L. D., 403, 433
Bessel's formula, for probable error of a single observation, 298; for probable error of the mean, 299; relation of constants in Peters' and Bessel's formulas, 303; tables to facilitate calculation of probable errors by, 305
Blank tests, analysis of field data from, 410; arrangement of, 409; meaning of, 409; to determine frequency or arrangement of check plots, 430; to determine necessary number of replications, 416
Borden, R. J., 442

Card system, in making a correlation table, 100; in making a frequency distribution, 17
Chaddock, R. E., 7, 8
Chang, C. C., 28
Check plots, 429; methods of obtaining calculated check yields from, 430; number of, 430
Chi-square test, application of, 287
Classes, limits of, 12; mid-points of, 12; number of, 11; range of, 11; recording data in, 15
Competition, plot, 420; effect of, in arranging plot layouts, 428; methods of measuring the effect of, 420
Contingency, coefficient of, 165; correction for, 169; limitations of, 169, 172; methods of determining, 165
Correlation, analysis of, 90; by foot-rule method, 164; from ranks, 161; perfect negative, 107; perfect positive, 105. *See* Multiple correlation, Non-linear correlation, Partial correlation, Probability
Correlation, coefficient of, 95; a measure of linear correlation, 95; a ratio value, 113, 116, 148; calculation of, 96; used for regression lines, 118
Correlation ratio, 147; comparison with correlation coefficient, 154; correction for, 155; method of calculating, 149
Correlation table, construction of, 98; importance of, 152
Covariance, analysis of, 442

Curved regression lines, 155; for prediction, 160; method of calculating, 155
Curves, fitting of, 249; criteria for, 249, 251, 252; examples of various types of curves and method of fitting, 252-283; groups of curve types, 252; logarithmic, 282

Data, collection of, 2; measuring and recording, 4
Day, James W., 407
Degrees of freedom, 289, 341
Deviation or dispersion, constants of, 66; comparison of usefulness of, 89; necessary qualifications of, 67; relation between average and standard deviations, 85; relation between standard and quartile deviations, 89; types of, 67. *See* Average deviation, Quartile deviation, Standard deviation; Variability
Distributions, types of frequency, 20

Elderton, W. Palin, 287, 288, 289, 290, 294, 296
Emerson, R. A., 292
Empirical mode, method of determining, 55
Errors of observation, 216
Ezekiel, Mordecai, 215

Field experiments, interpretation of results from, 452; for a period of years, 462; on basis of deviation from the mean method, 457; on basis of general probable error, 459; on basis of general probable error from check plots, 456; on basis of standard or probable error, 452
Fisher, R. A., 155, 240, 289, 290, 296, 331–335, 340, 341, 349, 350, 360, 442, 452, 462, 463
Fisher's method, for interpreting goodness of fit, 289; for interpreting results from small samples, 331; tabled values for argument *t*, 331
Foot-rule method, for determining correlation, 164
Fraser, A. C., 23
Frequency curves, 29; fitting of, 249
Frequency distributions, grouping of measurements in, 11; methods of making, 15; types of, 20; value of, 11

Garber, R. J., 403
Geometric mean, definition of, 41; method of calculating, 63; use of, 64
Goodness of Fit, in curve fitting, 287; in Mendelian segregation, 295; interpretation of, 289; method of determining, 287
Graphic illustration, importance of, 27; rules for, 36

Hall, A. D., 405
Harris, J. Arthur, 399, 401, 402
Hayes, H. K., 439, 457, 459
Histogram, 31
Hoover, M. M., 403

Illinois Agricultural Experiment Station, 62, 139, 140
Interpretation of results, general suggestions for, 7

Kiesselbach, T. A., 420

Latin square, analysis and arrangement of, 360
Least squares, 132
Line diagram, 27
Logarithmic curve, 282

McClelland, C. K., 406
McIlvaine, T. C., 403
Mean. *See* Arithmetic mean, Averages, Geometric mean
Median, definition of, 41; method of determining, 52. *See* Averages
Mills, Frederick C., 58
Missing plots, eliminating effect of, 364; methods of calculating yields of, 365
Mode, definition of, 41; types of, 55. *See* Empirical mode, Theoretical mode
Moving average, 61, 435
Multiple correlation, 173; coefficient of, 183; definition of, 173; limitations of method for, 215; method of determining, 173

Non-linear correlation, 147
Normal curve of error, 221, 272
Number of observations, to measure differences, 318

Odds, calculation of, 231; interpretation of, 236; tables of, 236, 317. *See* Probability
Ogive, 28
Ordinates, 29, 133; method of loaded or weighted, 30

Parabolas, methods of fitting second, third, and fourth order, 143
Parker, E. R., 403, 433
Partial correlation, 201; coefficient of, 202; definition of, 201; limitations of methods for, 215; method of determining, 202
Pearl, Raymond, 249
Pearson, K., 137, 148, 155, 165, 237, 251, 254, 258, 261, 262, 264, 273, 274, 287, 289, 296, 316, 335
Percentages, dangers in use of, 9
Peters' formula, for probable error of a single observation, 303; for probable error of the mean, 303; relation of constants in Bessel's and Peters' formulas, 303; tables to facilitate calculation of probable errors by, 305
Plot technic, 398; efficiency in use of land of plots of different sizes, 414; recommendation regarding plot arrangement and replication, 463; size, shape, and replication of plots, 404. *See* Blank tests, Check plots, Competition, Soil heterogeneity, Uniformity trials
Prediction, calculated from regression, 118, 125; from curved regression lines, 160; from straight lines, 139; Pearson's method of, based on past experience, 335; relation between predicted and actual values, 125; standard deviation of predicted values, 125, 179, 370
Probability, methods of interpreting, 231; for results from small samples, 324
Probable error, concept of, 216; determined by counting, 230;

formulas for probable errors of various constants, 244; of an average, 246; of a single observation, 224; of a sum or difference, 243, 305, 307; of Mendelian results, 337; of the coefficient of correlation, 239; of the coefficient of variability, 239; of the correlation ratio, 240; of the difference between the correlation coefficient and correlation ratio, 242; of the mean, 225; of the probable error of a single determination, 248; of the standard deviation, 238; to analyze results from several trials, 311. *See* Bessel's formula, Peters' formula, Probability

Quartile deviation, 88, 230

Random arrangement, 342; restriction of, 359
Random sample, 3
Rank correlation, 161
Regression, application of, 122; based on three variables, 186; based on two variables, 183
Regression lines, 118; method of fitting, 122. *See* Curved regression lines, Variance analysis
Relation, between two characters, 90; by means of fitted lines, 132; by means of graphs, 93. *See* Correlation
Richey, Frederick D., 435, 437, 438, 439, 441
Russell, E. J., 405

Sanders, H. G., 442
Scatter diagram, 91
Scofield, C. S., 402
Secrist, Horace, 64
Shen, T. H., 423, 424, 425
Sheppard's correction, 250, 276
Skewness, 251
Small samples, methods for evaluating results obtained from, 297. *See* Bessel's formula, Fisher's method, Peters' formula, Probability, 'Student's' method
Snedecor, George W., 180, 190, 240, 331, 350, 452
Soil heterogeneity, 399
Spearman, C., 161, 164
Stadler, L. J., 420, 422, 426, 427
Standard deviation, definition of, 74; methods of determining, 74; of a series of standard deviations, 86. *See* Standard error of estimate
Standard error, 223; in variance analysis, 340; relation between standard and probable error, 224
Standard error of estimate, in correlation analysis, 129; in multiple correlation, 179; in variance analysis, 369
Standards, for calculation of results, 6
Straight line, for prediction, 139; method of fitting, 132
Stringfield, G. H., 423, 424, 425
'Student,' 10, 324–331, 335, 425, 459, 462, 463
'Student's' method, for interpreting results from small samples, 324; for argument *t*, 328; table of odds for argument *t*, 328; table of odds for argument *Z*, 325

Tables for aid in calculation and interpretation; corresponding values for r computed from r_r, 472; for estimating probability based on the normal probability integral corresponding to values of x/σ, 475; odds calculated for the Z values of 'Student's' table, 480; odds calculated from 'Student's' t table, 484, 485; odds for differences in one direction only, 317; odds for values of $D/P.E.$, 479; odds for values of $D/P.E.$ (short table), 236; significant values of r and R, 490; sums of powers of natural numbers, 468; values for interpreting Goodness of Fit, 291; values of F and t, 486; values to facilitate computation of the probable error of a single observation, from Bessel's formula, 473; values to facilitate computation of the probable error of a single observation, from Peters' formula, 474; values to facilitate computation of the probable error of the mean, from Bessel's formula, 473; values to facilitate computation of the probable error of the mean, from Peters' formula, 474; values to facilitate fitting of a logarithmic curve, 469
Theoretical mode, 55; approximate, 56; comparisons of approximate and true, 57; methods of determining, 56. *See* Averages
Tippett, L. H. C., 155

Uniformity trials, 441
United States Department of Agriculture Weather Bureau, 24

Variability, coefficient of, 86
Variance, 67, 74
Variance analysis, 340; application of method of, 342; applied to a complex experiment, 378; applied to problems of plot technic, 412; applied to regression analysis, 369; applied to results obtained by fitting curved regression lines, 375; interpretation from, 348

Wallace, H. A., 180, 190
Wishart, J., 364, 365, 452

Yates, F., 364, 366
Yule, G. Udny, 21, 40, 41, 57, 64, 74, 119, 168, 169